2018 EE448, Big Data Mining, Lecture 12

# Real-Time Bidding & Behavioral Targeting

Weinan Zhang

Shanghai Jiao Tong University

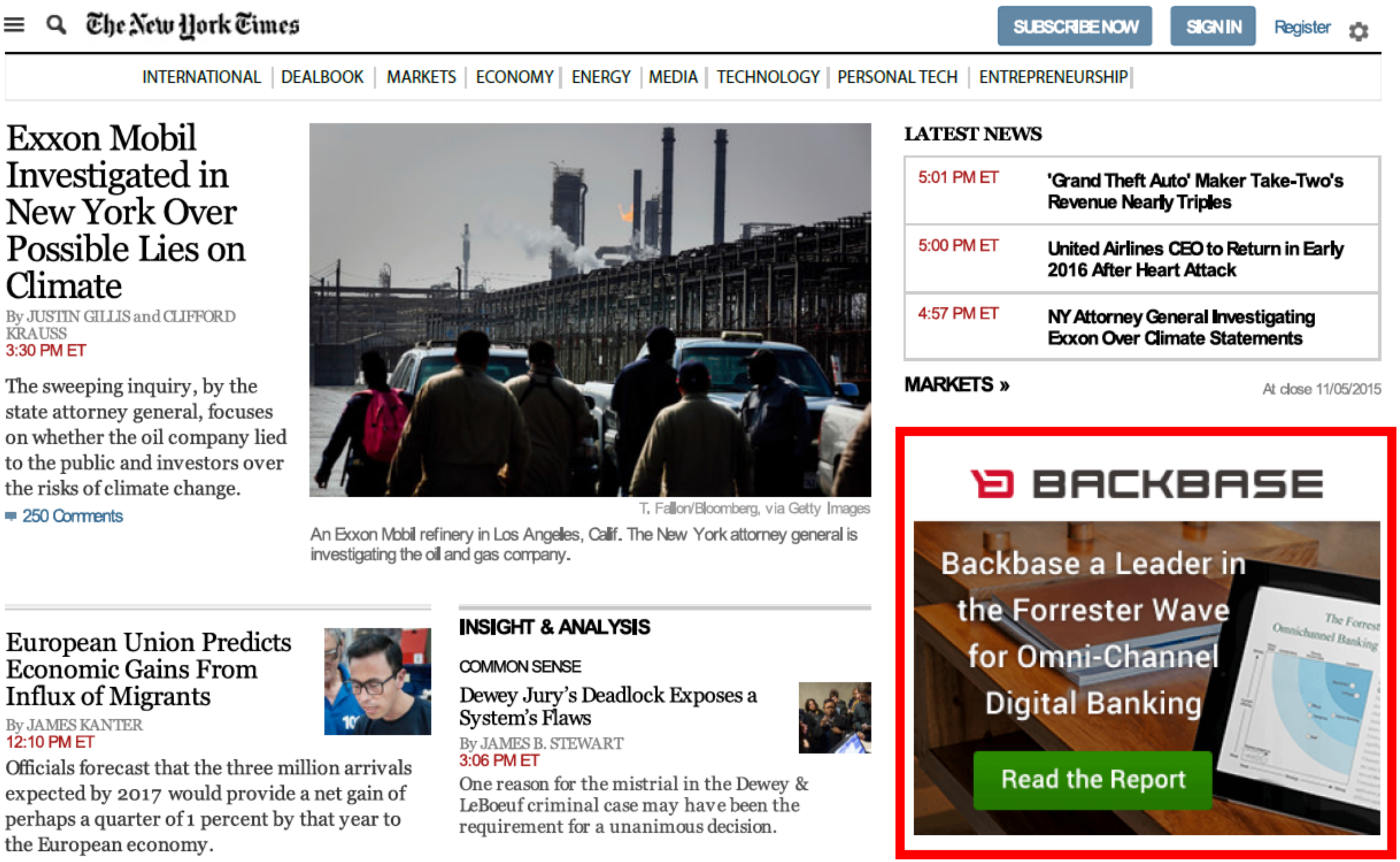
http://wnzhang.net

http://wnzhang.net/teaching/ee448/index.html

# Content of This Course

- Real-time bidding based display advertising
- User tracking and profiling
- Real-time bidding strategies
- Fraud detection

# Display Advertising



The New York Times
SUBSCRIBE NOW | SIGN IN | Register

INTERNATIONAL | DEALBOOK | MARKETS | ECONOMY | ENERGY | MEDIA | TECHNOLOGY | PERSONAL TECH | ENTREPRENEURSHIP

## Exxon Mobil Investigated in New York Over Possible Lies on Climate

By JUSTIN GILLIS and CLIFFORD KRAUSS
3:30 PM ET

The sweeping inquiry, by the state attorney general, focuses on whether the oil company lied to the public and investors over the risks of climate change.

250 Comments

T. Fallon/Bloomberg, via Getty Images

An Exxon Mobil refinery in Los Angeles, Calif. The New York attorney general is investigating the oil and gas company.

**LATEST NEWS**

| | |
|---|---|
| 5:01 PM ET | 'Grand Theft Auto' Maker Take-Two's Revenue Nearly Triples |
| 5:00 PM ET | United Airlines CEO to Return in Early 2016 After Heart Attack |
| 4:57 PM ET | NY Attorney General Investigating Exxon Over Climate Statements |

**MARKETS »** At close 11/05/2015



## European Union Predicts Economic Gains From Influx of Migrants

By JAMES KANTER
12:10 PM ET

Officials forecast that the three million arrivals expected by 2017 would provide a net gain of perhaps a quarter of 1 percent by that year to the European economy.

**INSIGHT & ANALYSIS**

COMMON SENSE

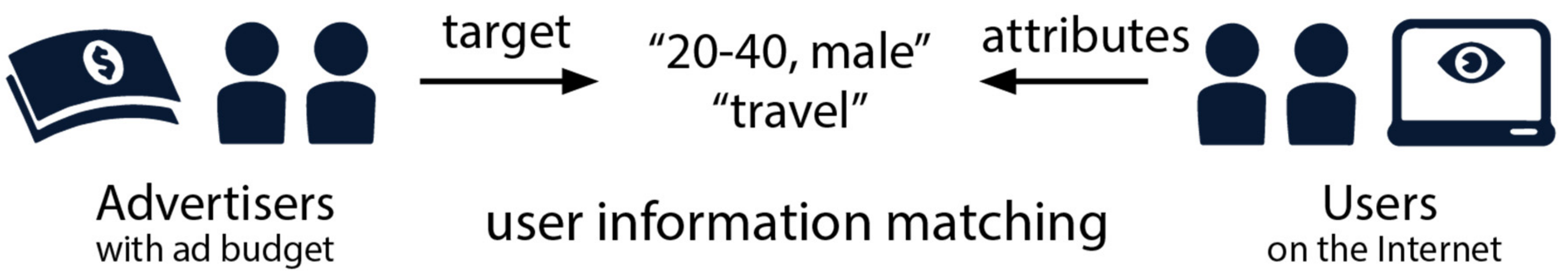
## Dewey Jury's Deadlock Exposes a System's Flaws

By JAMES B. STEWART
3:06 PM ET

One reason for the mistrial in the Dewey & LeBoeuf criminal case may have been the requirement for a unanimous decision.

http://www.nytimes.com/

# Display Advertising




- Advertiser targets a segment of users
  - No matter what the user is searching or reading
- Intermediary matches users and ads by user information

# Internet Advertising Frontier:

## Real-Time Bidding (RTB) based Display Advertising

### What is Real-Time Bidding?

- Every online ad view can be evaluated, bought, and sold, all individually, and all instantaneously.
- Instead of buying keywords or a bundle of ad views, advertisers are now buying users directly.

- Behavioral targeting: it is possible now to track user actions resulted from an online campaign, advertising optimization becomes more resembling to that of the financial market trading and tends to be driven by the marketing profit and return-on-investment (ROI).

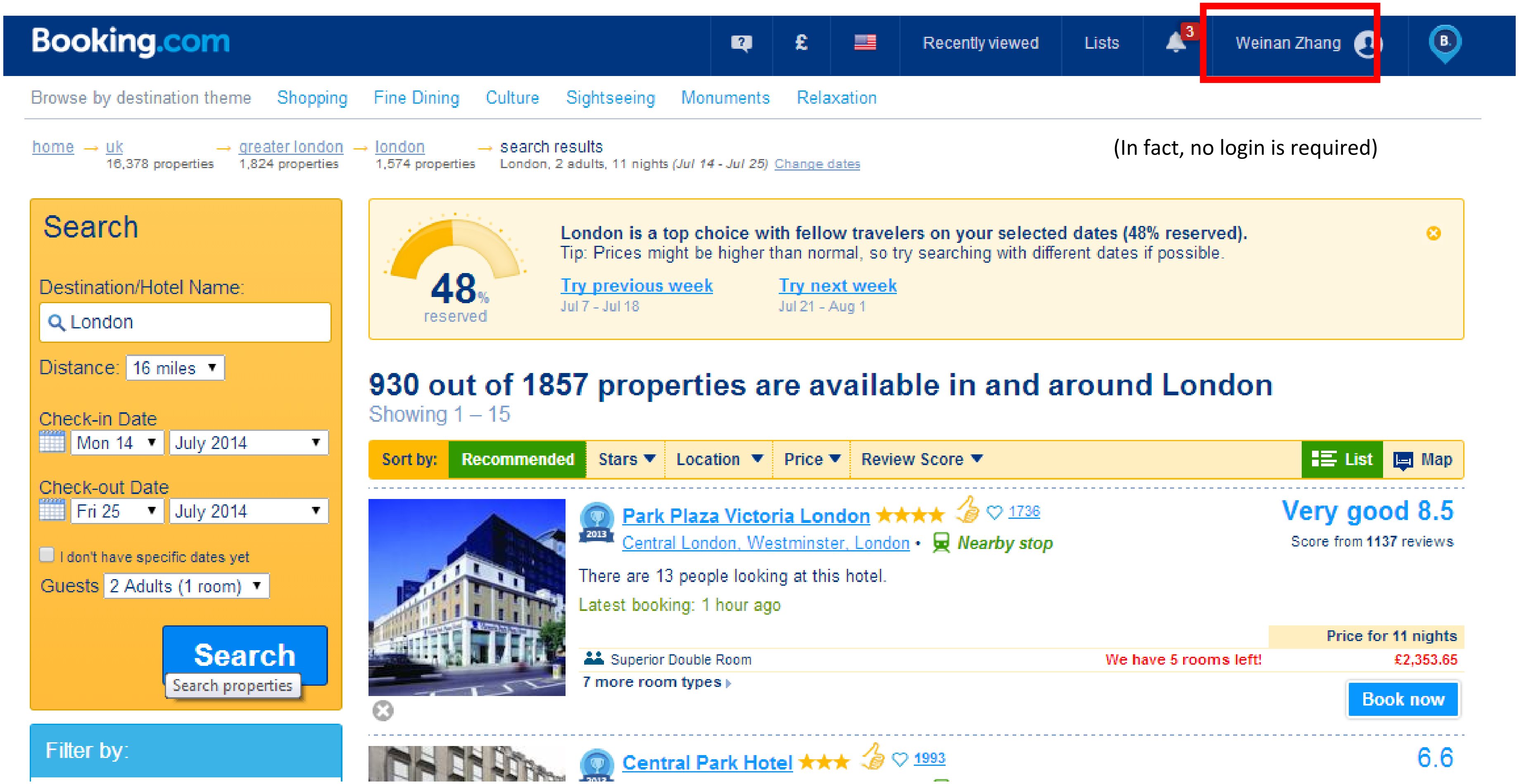
# An Example of RTB

Suppose a student regularly reads articles on emarketer.com


Content-related ads

# An Example of RTB

## He recently checked the London hotels



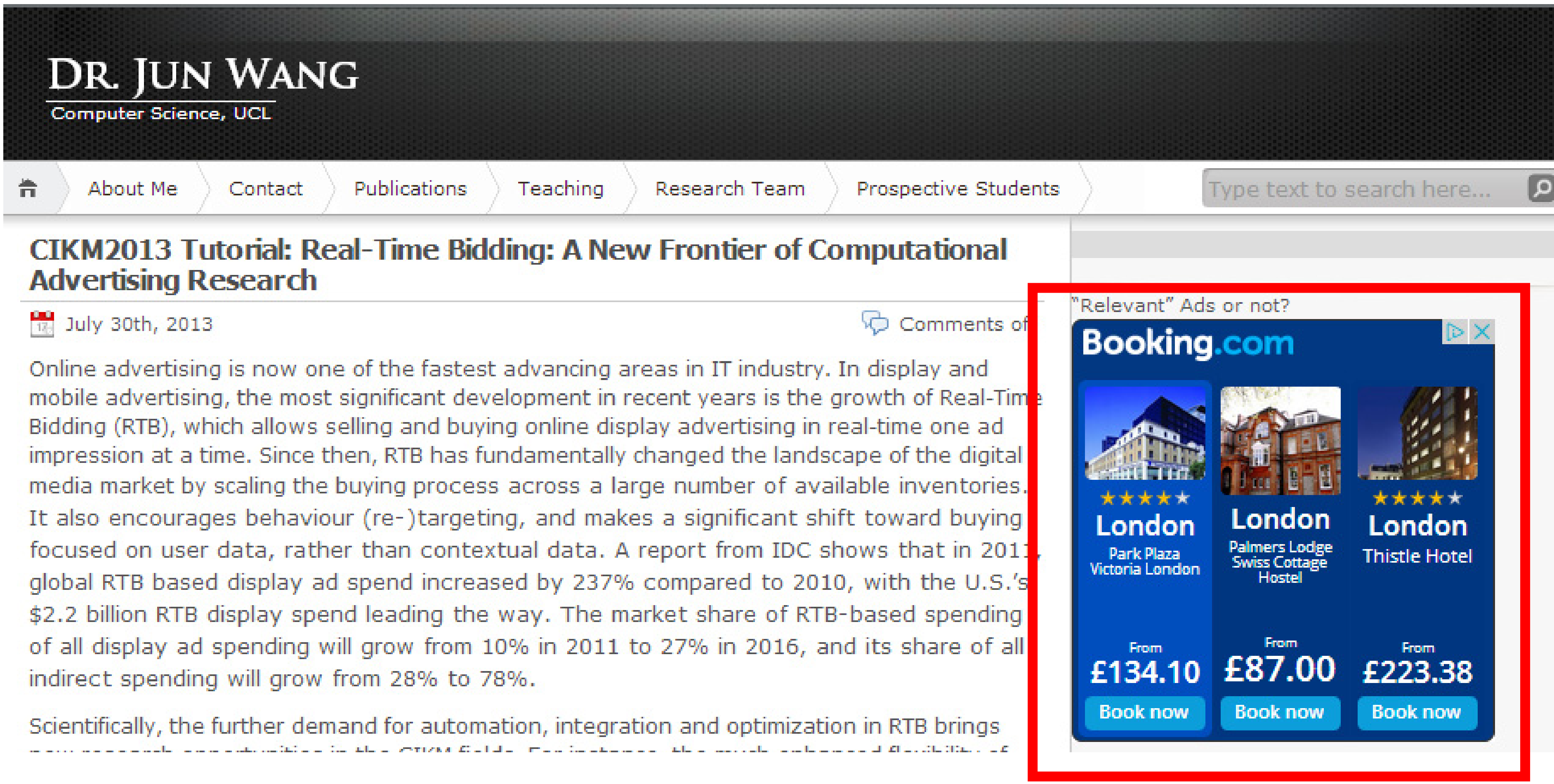

(In fact, no login is required)

# An Example of RTB

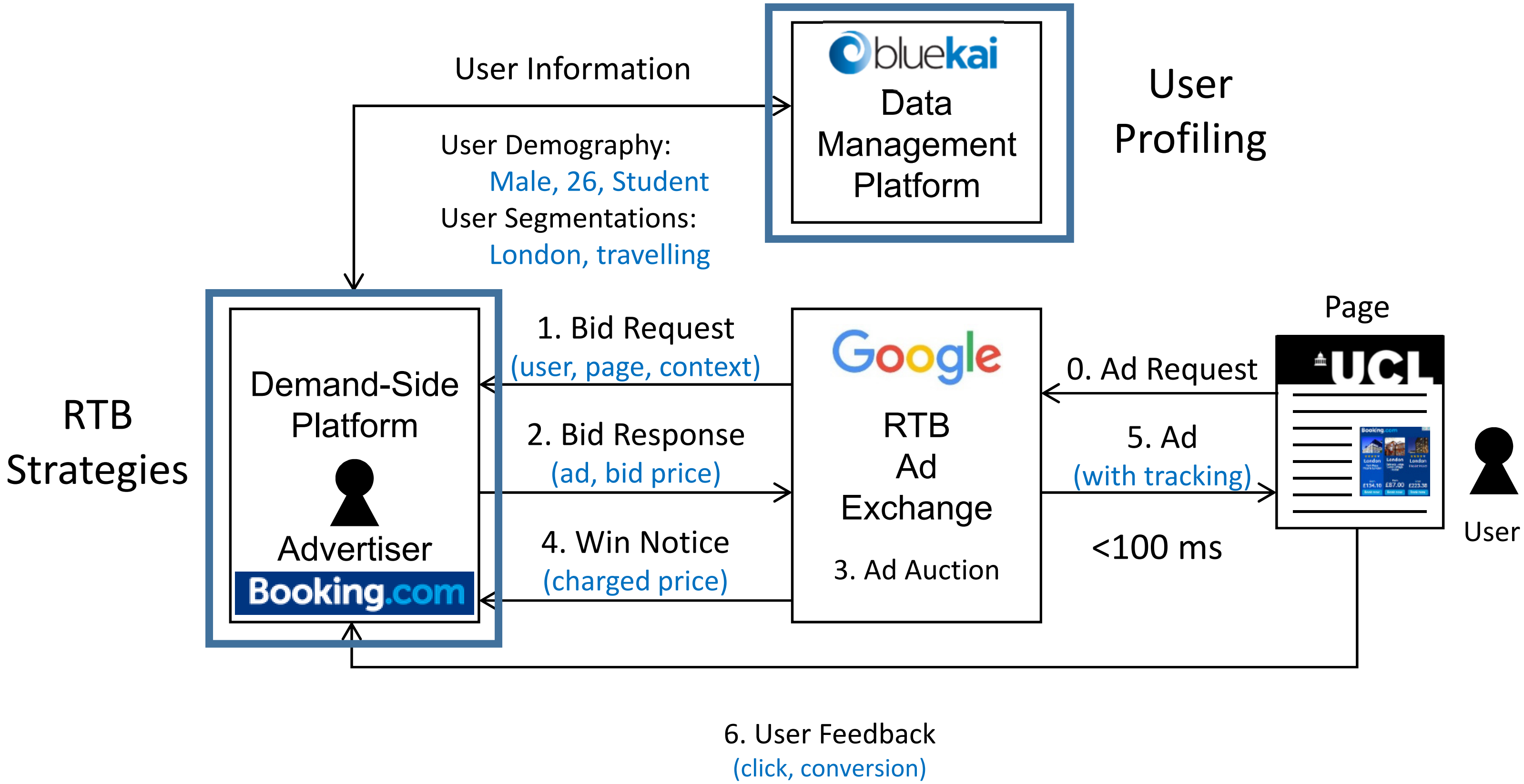
## Relevant ads on facebook.com

# An Example of RTB

Even on supervisor’s homepage!
(User targeting dominates the context)



# RTB Display Advertising Mechanism




- Buying ads via real-time bidding (RTB), 10 billion per day
- A real big data battlefield

# RTB: A Big Data Battle Field

- The daily volume of RTB platforms and the comparison with finance institutes

| | DSP/Exchange | Daily Traffic |
|---|---|---|
| Advertising | iPinYou, China | 18 billion impressions |
| | YOYI, China | 5 billion impressions |
| | Fikisu, US | 32 billon impressions |
| Finance | New York Stock Exchange | 12 billion shares |
| | Shanghai Stock Exchange | 14 billion shares |

| | Query per Second |
|---|---|
| Turn DSP | 1.6 million |
| Google | 40,000 search |

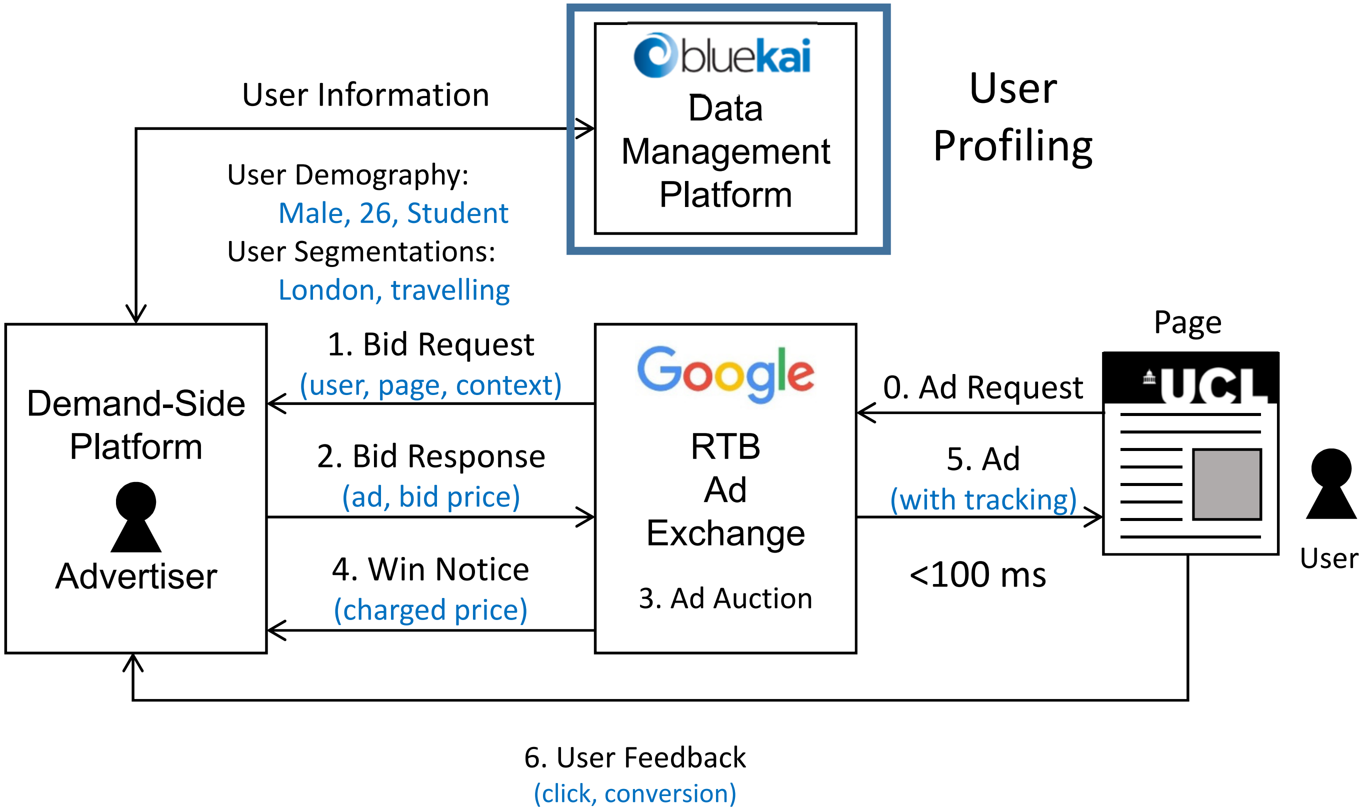
It is fair to say that the transaction volume from display advertising has already surpassed that of the financial market

Zhang, Haifeng, Zhang, Weinan et al. "Managing Risk of Bidding in Display Advertising“. WSDM 2017.
Shen, Jianqiang, et al. "From 0.5 Million to 2.5 Million: Efficiently Scaling up Real-Time Bidding." ICDM 2015.

# Content of This Course

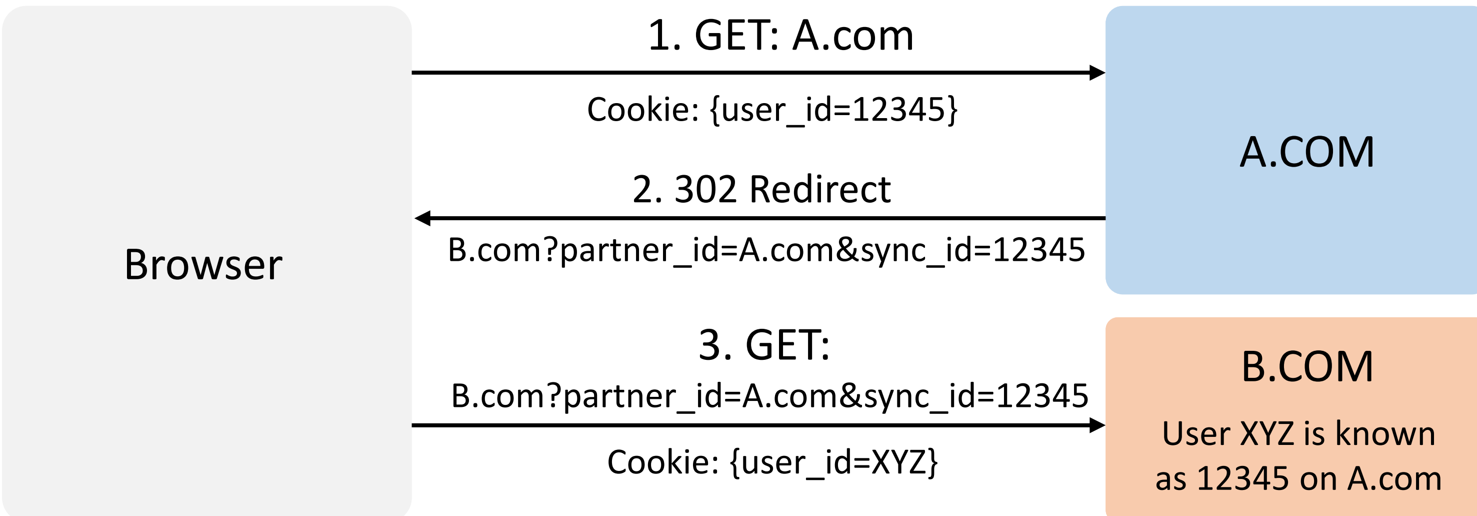
- Real-time bidding based display advertising
- User tracking and profiling
- Real-time bidding strategies
- Fraud detection

# DMP: Data Management Platform




- DMP is a data warehouse that stores, merges, and sorts, and labels it out in a way that's useful for marketers, publishers and other businesses.

# Cookie Sync: Merging Audience Data




When a user visits a site (e.g. ABC.com) including A.com as a third-party tracker.

(1) The browser makes a request to A.com, and included in this request is the tracking cookie set by A.com.

(2) A.com retrieves its tracking ID from the cookie, and redirects the browser to B.com, encoding the tracking ID into the URL.

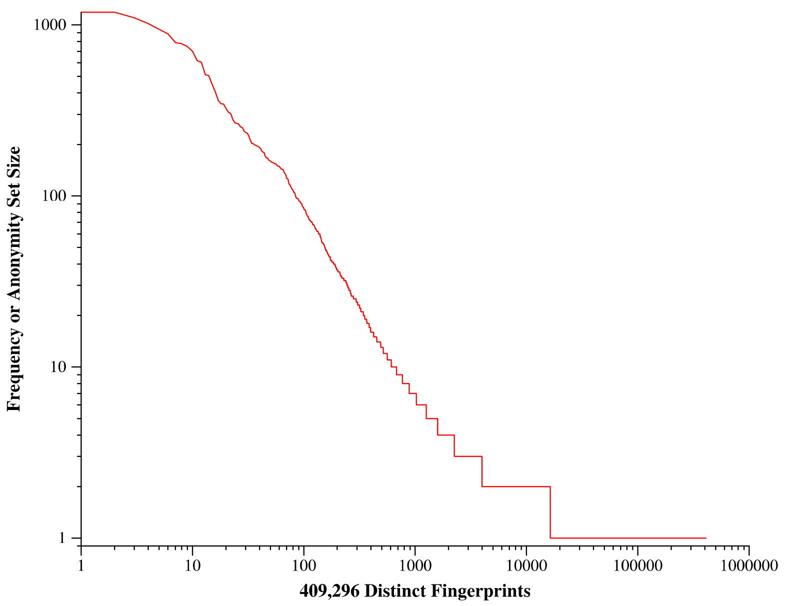
(3) The browser then makes a request to B.com, which includes the full URL A.com redirected to as well as B.com's tracking cookie.

(4) B.com can then link its ID for the user to A.com's ID for the user2

https://freedom-to-tinker.com/blog/englehardt/the-hidden-perils-of-cookie-syncing/

# Browser Fingerprinting

- A device fingerprint or browser fingerprint is information collected about the remote computing device for the purpose of identifying the user.
- Fingerprints can be used to fully or partially identify individual users or devices even when cookies are turned off.

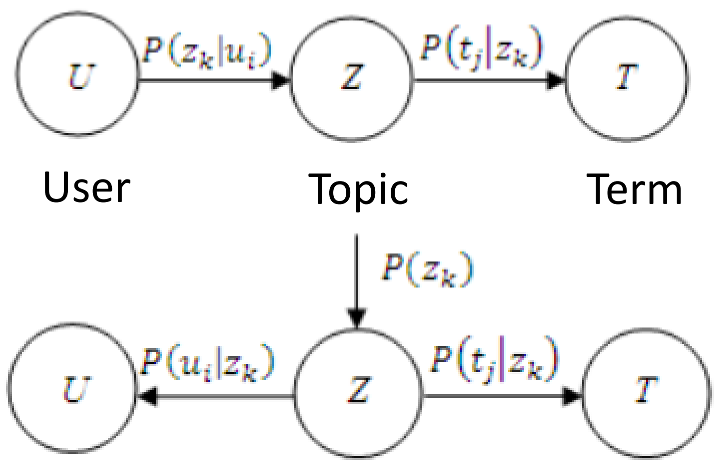



94.2% of browsers with Flash or Java were unique in a study

Eckersley, Peter. "How unique is your web browser?." Privacy Enhancing Technologies. Springer Berlin Heidelberg, 2010.
Acar, Gunes, et al. "The web never forgets: Persistent tracking mechanisms in the wild." Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2014.

# User Segmentation and Behavioral Targeting

- Behavioral targeting helps online advertising
- From user – documents to user – topics
  - Latent Semantic Analysis / Latent Dirichlet Allocation

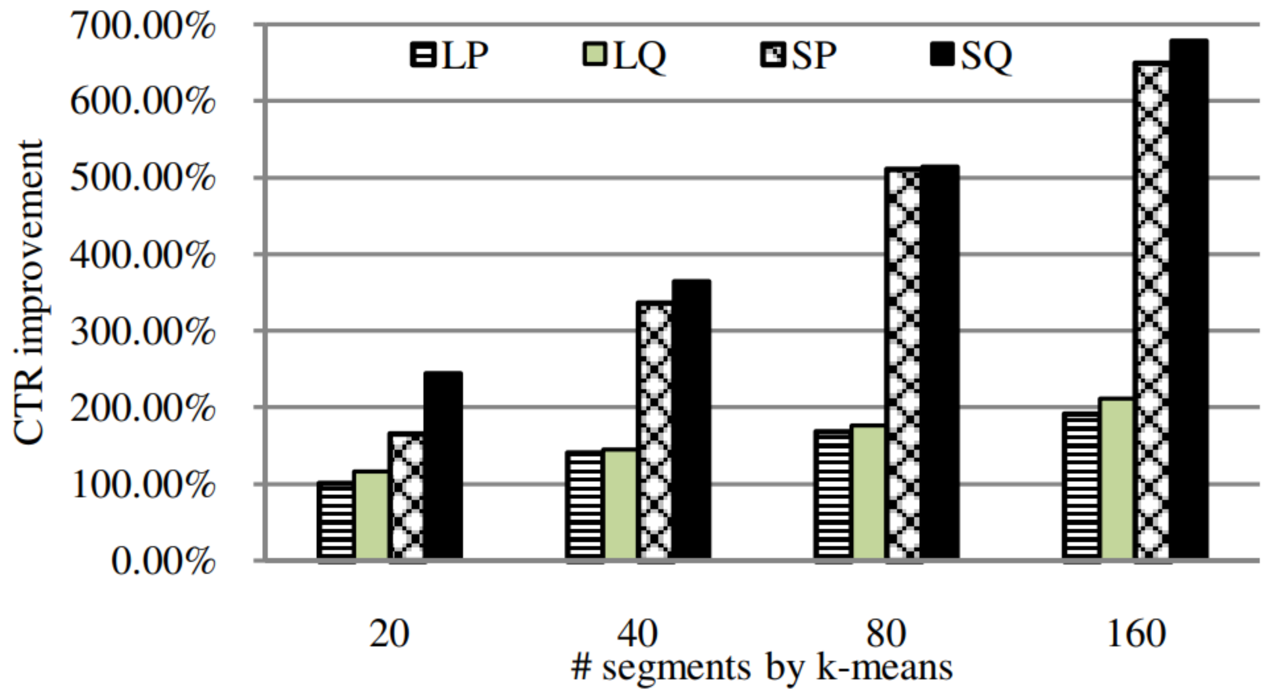




J Yan, et al., How much can behavioral targeting help online advertising? WWW 2009

X Wu, et al., Probabilistic latent semantic user segmentation for behavioral targeted advertising, Intelligence for Advertising 2009

# User Segmentation and Behavioral Targeting

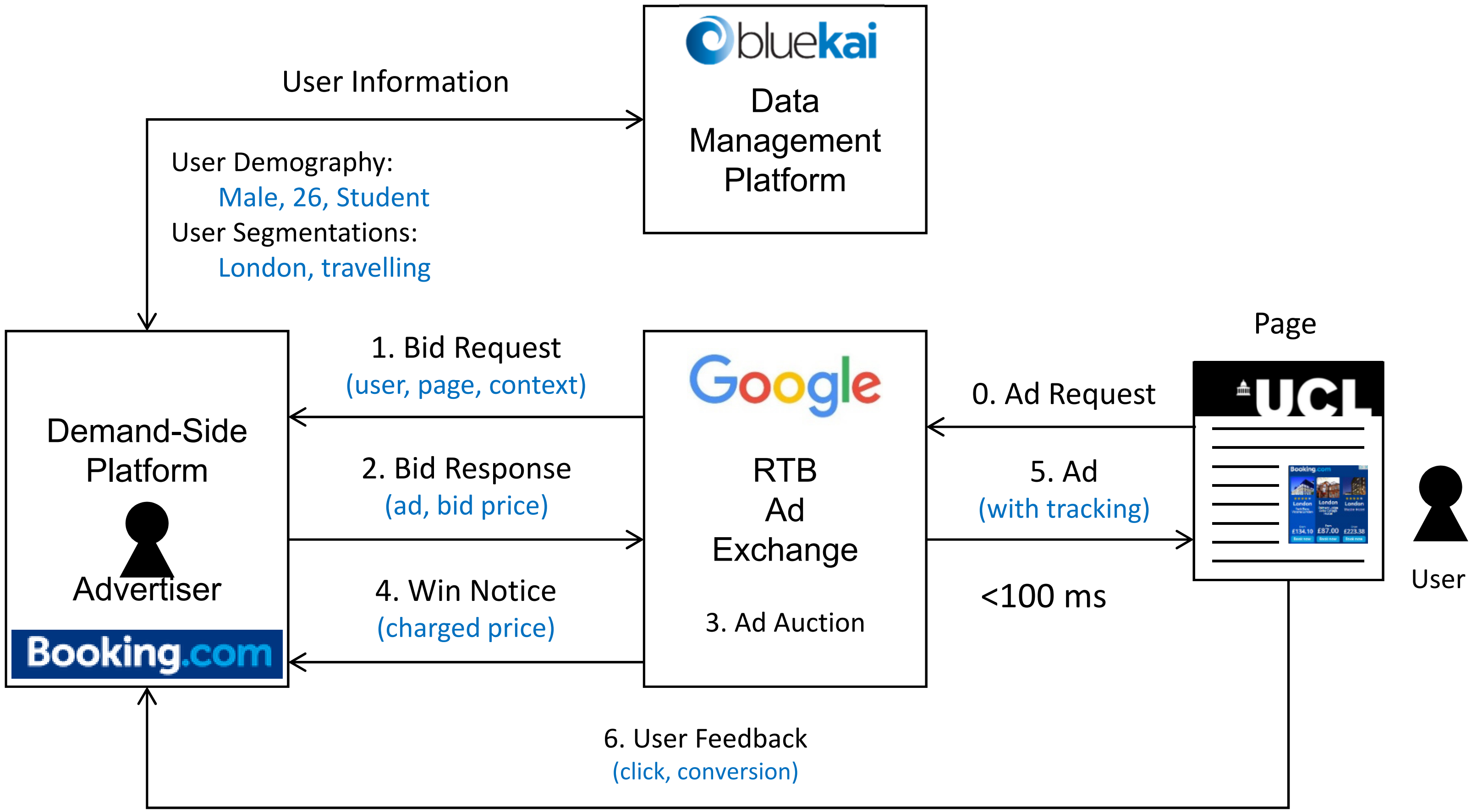



- LP: using Long term 7-day user behavior and representing the user behavior by Page-views;
- LQ: using Long term 7-day user behavior and representing the user behavior by Query terms;
- SP: using Short term 1-day user behavior and representing user behavior by Page-views;
- SQ: using Short term 1-day user behavior and representing user behavior by Query terms.

# Content of This Course

- Real-time bidding based display advertising
- User tracking and profiling
- Real-time bidding strategies
- Fraud detection

# RTB Display Advertising Mechanism




- Buying ads via real-time bidding (RTB), 10B per day

# Data of Learning to Bid

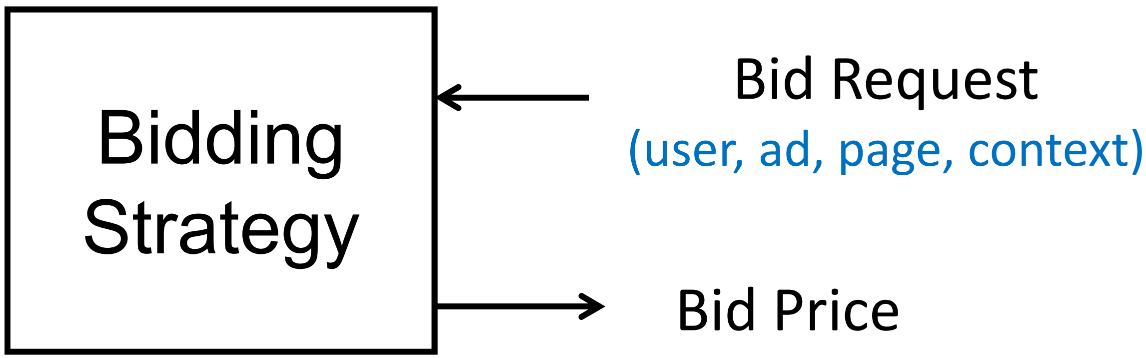
- Data

| $(\mathbf{x}, t)$ | $b$ | $w$ | $c$ | $y$ |
|---|---|---|---|---|
| (up,1500×20,Shanghai,0) | 5 | 1 | 4 | 1 |
| (down,1200×25,Paris,1) | 4 | 1 | 3 | 0 |
| (left,20×1000,Los Angeles,2) | 3 | 0 | × | × |
| (right,35×600,London,3) | 0 | 0 | × | × |

- Bid request features: High dimensional sparse binary vector
- Bid: Non-negative real or integer value
- Win: Boolean
- Cost: Non-negative real or integer value
- Feedback: Binary

# Problem Definition of Learning to Bid

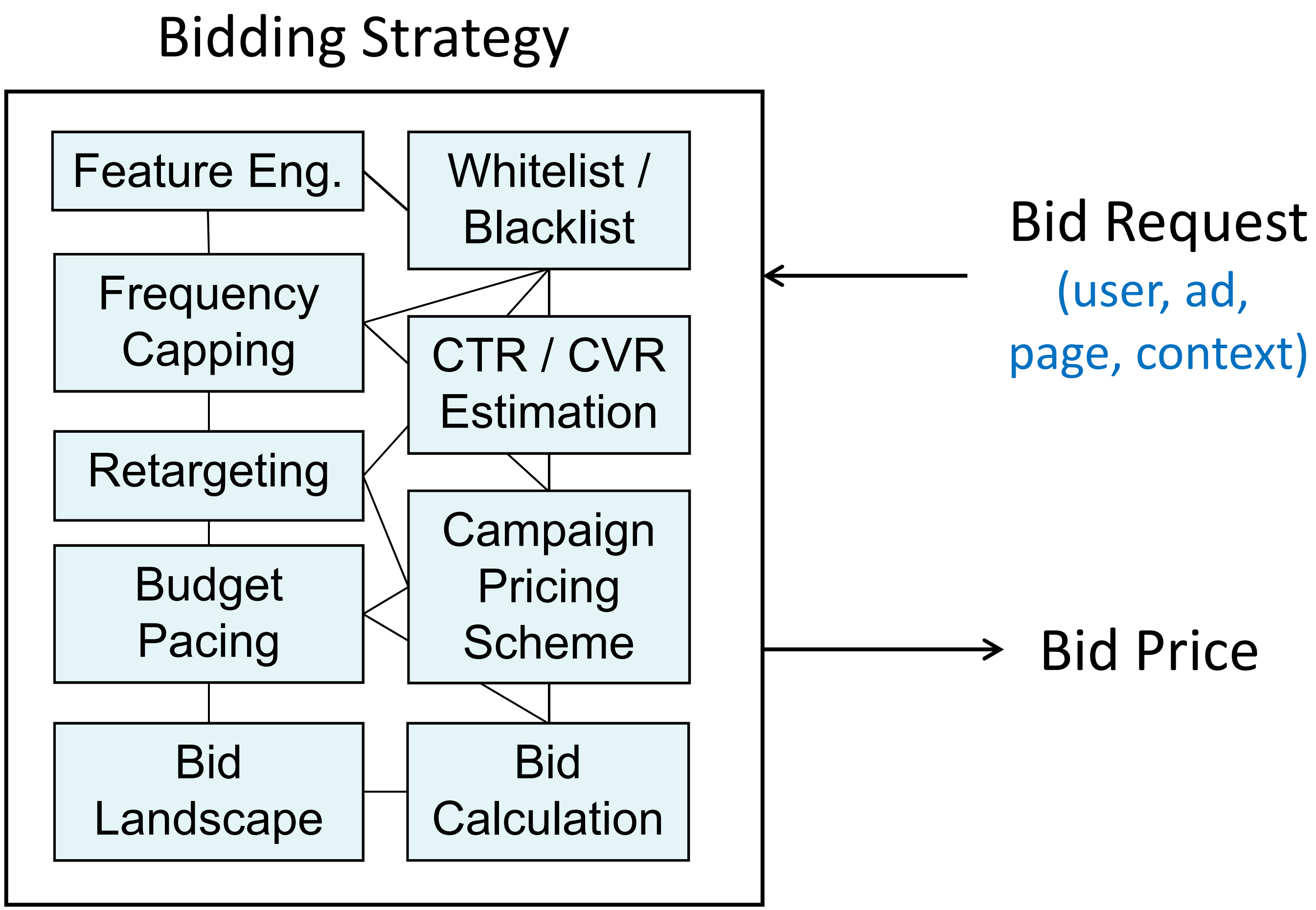
- How much to bid for each bid request?
  - Find an optimal bidding function b(x)



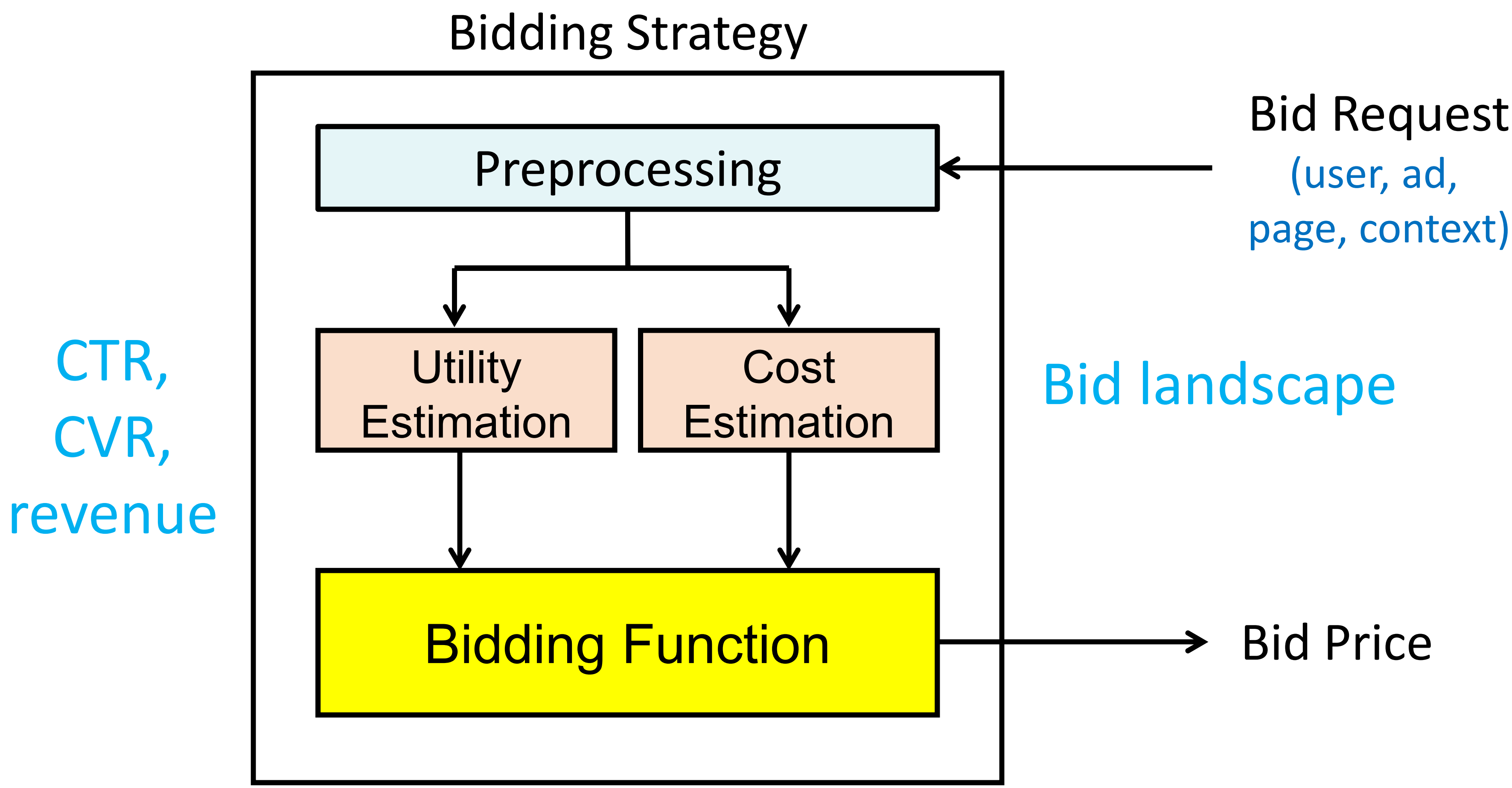

- Bid to optimize the KPI with budget constraint

$$\max_{\text{bidding strategy}} \quad \text{KPI}$$
$$\text{subject to} \quad \text{cost} \leq \text{budget}$$

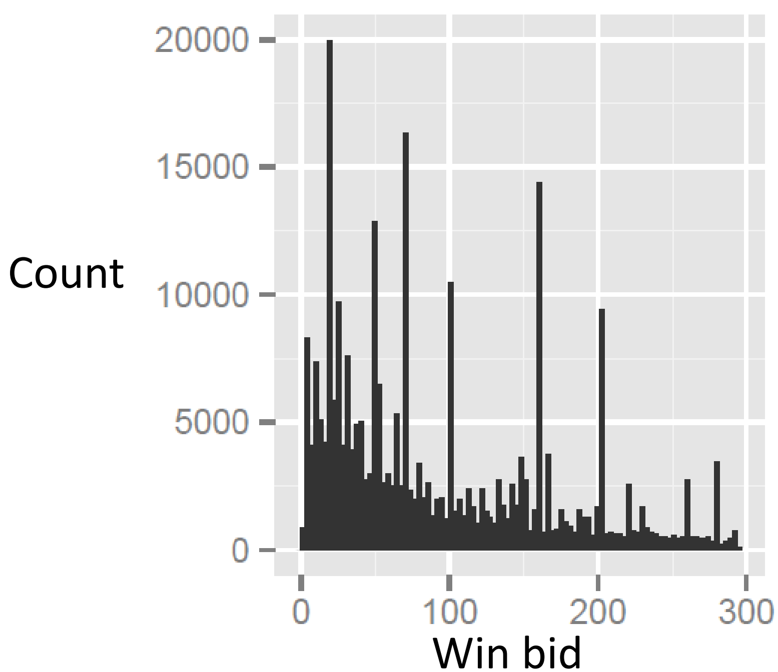
# Bidding Strategy in Practice

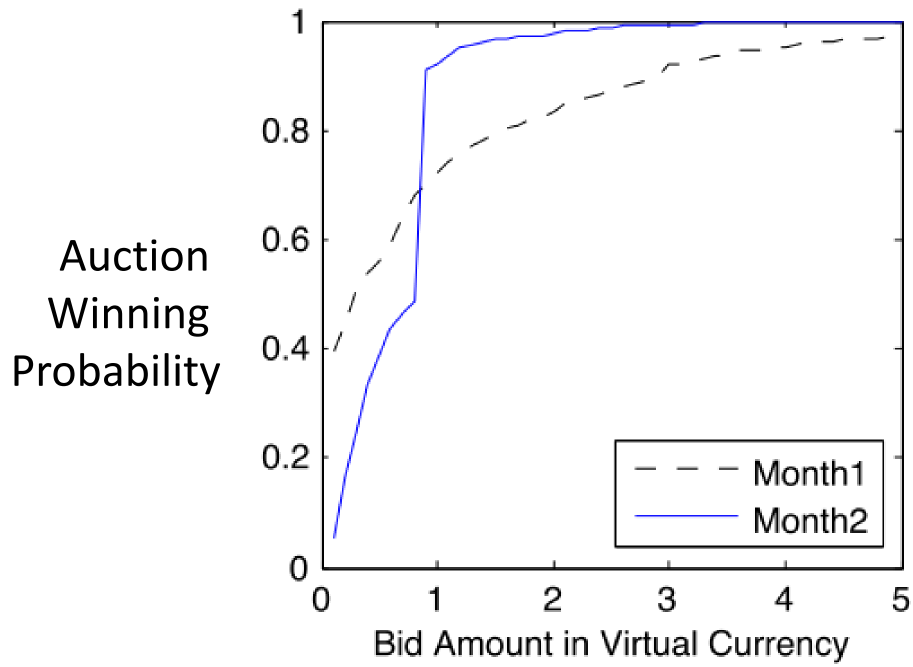


# Bidding Strategy in Practice: A Quantitative Perspective



# Bid Landscape Forecasting



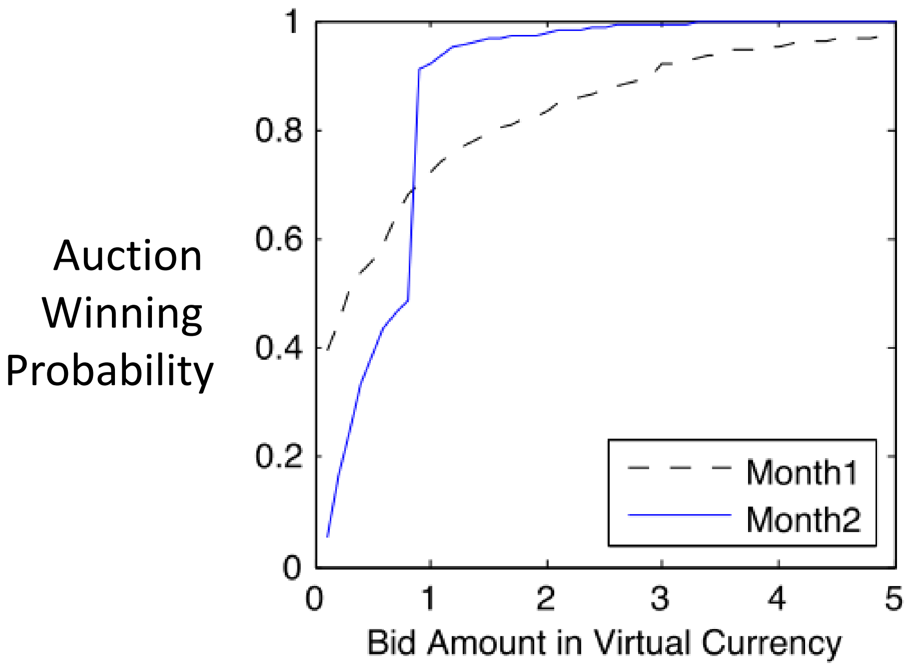

Win probability:

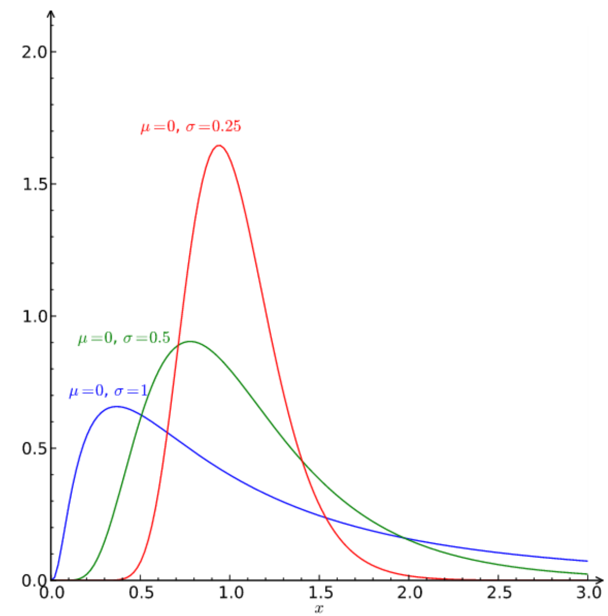
$$w(b) = \int_{z=0}^{b} p(z)dz$$

Expected cost:

$$c(b) = \frac{\int_{z=0}^{b} zp(z)dz}{\int_{z=0}^{b} p(z)dz}$$

# Bid Landscape Forecasting




- Log-Normal Distribution

$$f_{\mathbf{s}}(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}e^{\frac{-(ln x-\mu)^2}{2\sigma^2}},x>0$$

[Cui et al. Bid Landscape Forecasting in Online Ad Exchange Marketplace. KDD 11]

# Data Bias Problem for Bid Landscape

$$w(b) = \int_{z=0}^{b} p(z)dz$$

- If we directly count the probability from observed market prices

$$w_o(b_{\boldsymbol{x}}) = \frac{\sum_{(\boldsymbol{x}',y,z)\in D} \delta(z < b_{\boldsymbol{x}})}{|D|}$$

- The estimation is unbiased since the observed market prices is always lower than the historic bid
- Counterfactual case: example of WW2 planes

# Survival Model for Bid Landscape

- Kaplan-Meier Product-Limit method

$$l(b_{\boldsymbol{x}}) = \prod_{b_j < b_{\boldsymbol{x}}} \frac{n_j - d_j}{n_j} \qquad w(b_{\boldsymbol{x}}) = 1 - \prod_{b_j < b_{\boldsymbol{x}}} \frac{n_j - d_j}{n_j}$$

| $b_i$ | $w_i$ | $z_i$ |
|---|---|---|
| 2 | win | 1 |
| 3 | win | 2 |
| 2 | lose | $\times$ |
| 3 | win | 1 |
| 3 | lose | $\times$ |
| 4 | lose | $\times$ |
| 4 | win | 3 |
| 1 | lose | $\times$ |

| $b_j$ | $n_j$ | $d_j$ | $\frac{n_j - d_j}{n_j}$ | $w(b_j)$ | $w_o(b_j)$ |
|---|---|---|---|---|---|
| 1 | 8 | 0 | 1 | $1 - 1 = 0$ | 0 |
| 2 | 7 | 2 | $\frac{5}{7}$ | $1 - \frac{5}{7} = \frac{2}{7}$ | $\frac{2}{4}$ |
| 3 | 4 | 1 | $\frac{3}{4}$ | $1 - \frac{5}{7}\frac{3}{4} = \frac{13}{28}$ | $\frac{3}{4}$ |
| 4 | 2 | 1 | $\frac{1}{2}$ | $1 - \frac{5}{7}\frac{3}{4}\frac{1}{2} = \frac{41}{56}$ | $\frac{4}{4}$ |

# Survival Model for Bid Landscape

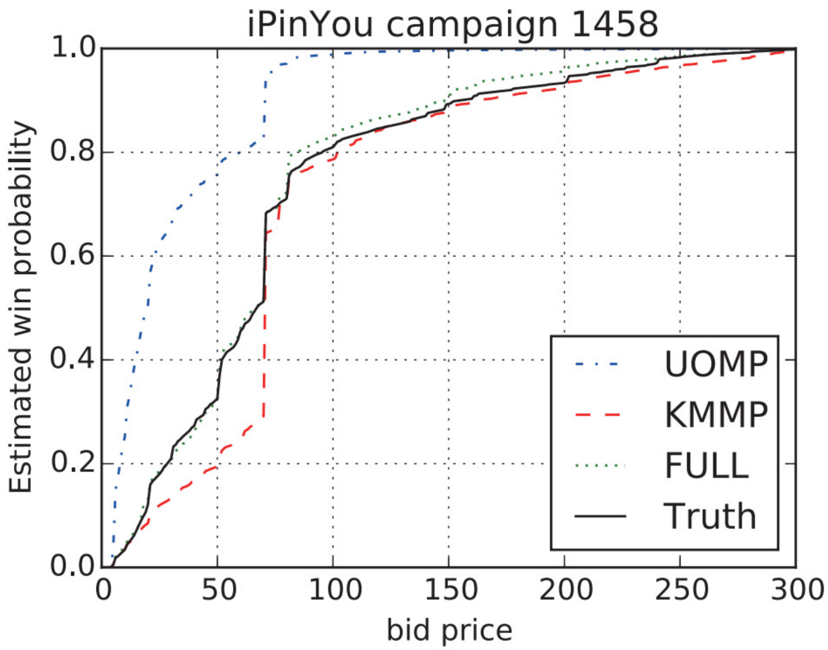
- Kaplan-Meier Product-Limit method

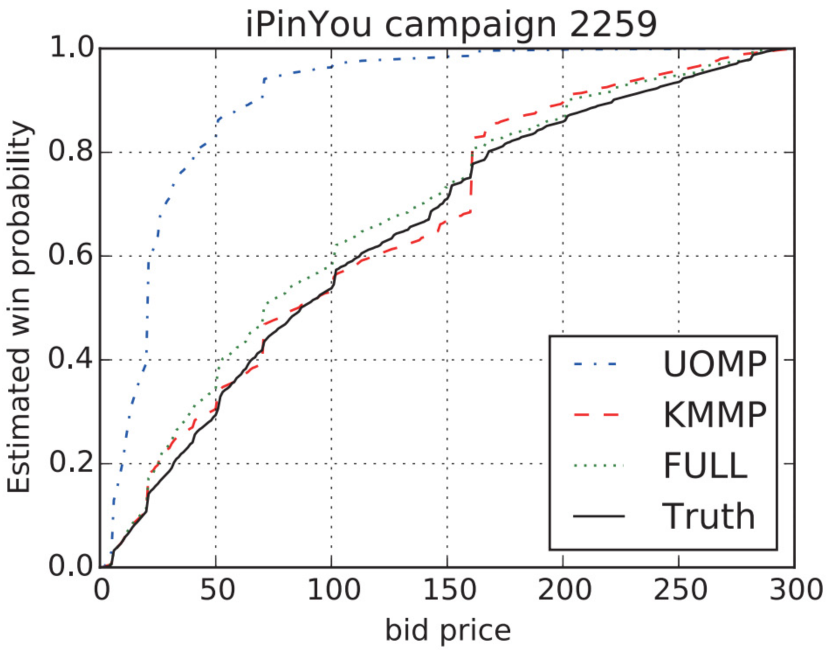
$$w_o(b_{\boldsymbol{x}}) = \frac{\sum_{(\boldsymbol{x}', y, z) \in D} \delta(z < b_{\boldsymbol{x}})}{|D|}$$

UOMP

$$w(b_{\boldsymbol{x}}) = 1 - \prod_{b_j < b_{\boldsymbol{x}}} \frac{n_j - d_j}{n_j}$$

KMMP



# Bid Landscape Forecasting

- Price Prediction via Linear Regression

$$z = \boldsymbol{\beta}^T \boldsymbol{x} + \epsilon \qquad \max_{\boldsymbol{\beta}} \sum_{i \in W} \log \phi\Big(\frac{z_i - \boldsymbol{\beta}^T \boldsymbol{x}_i}{\sigma}\Big)$$

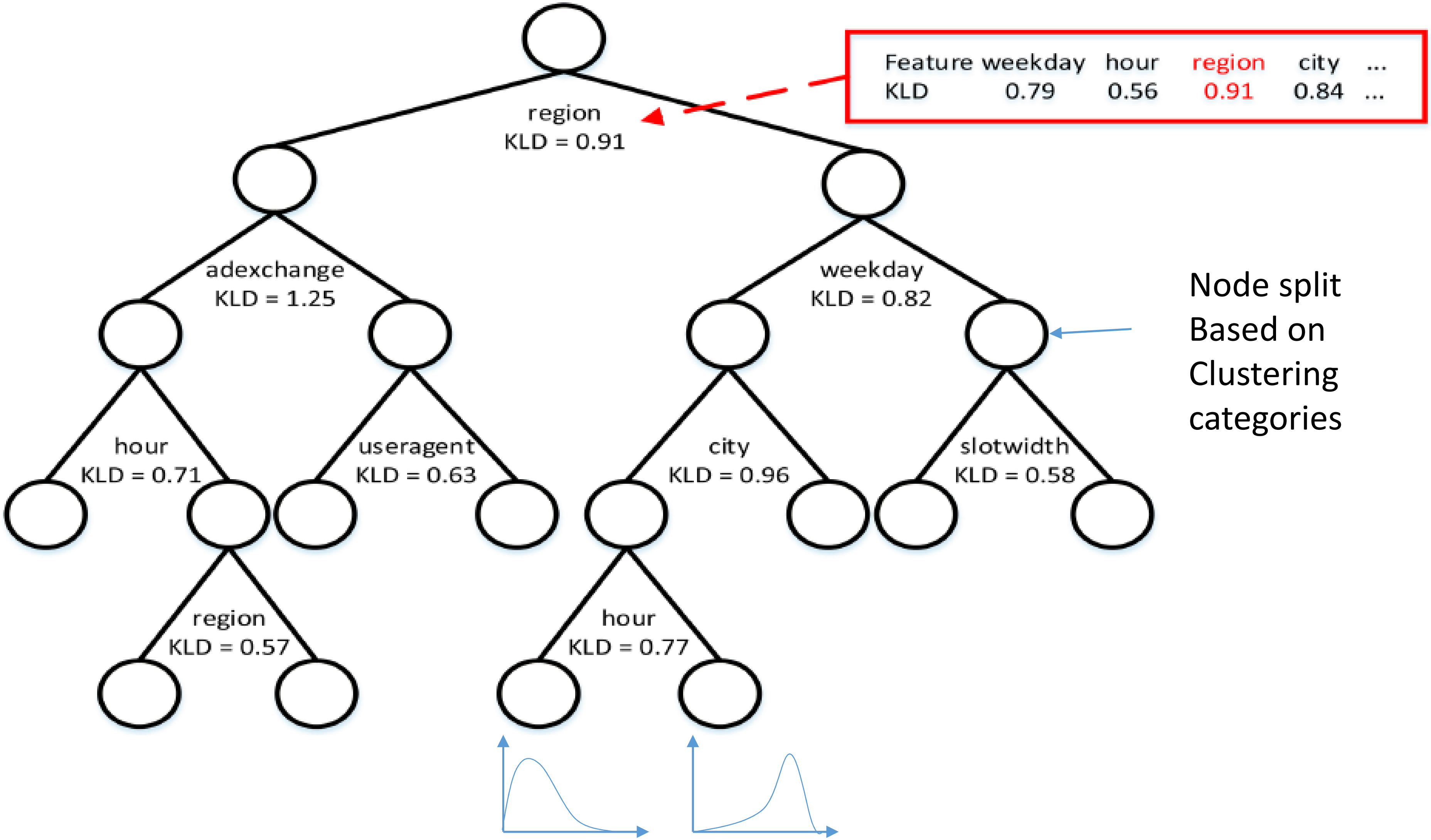
  - Modelling censored data in lost bid requests

$$P(b_i < z_i) = \Phi\Big(\frac{\boldsymbol{\beta}^T \boldsymbol{x}_i - b_i}{\sigma}\Big)$$

$$\max_{\boldsymbol{\beta}} \sum_{i \in W} \log \phi\Big(\frac{z_i - \boldsymbol{\beta}^T \boldsymbol{x}_i}{\sigma}\Big) + \sum_{i \in L} \log \Phi\Big(\frac{\boldsymbol{\beta}^T \boldsymbol{x}_i - b_i}{\sigma}\Big)$$

[Wu et al. Predicting Winning Price in Real Time Bidding with Censored Data. KDD 15]

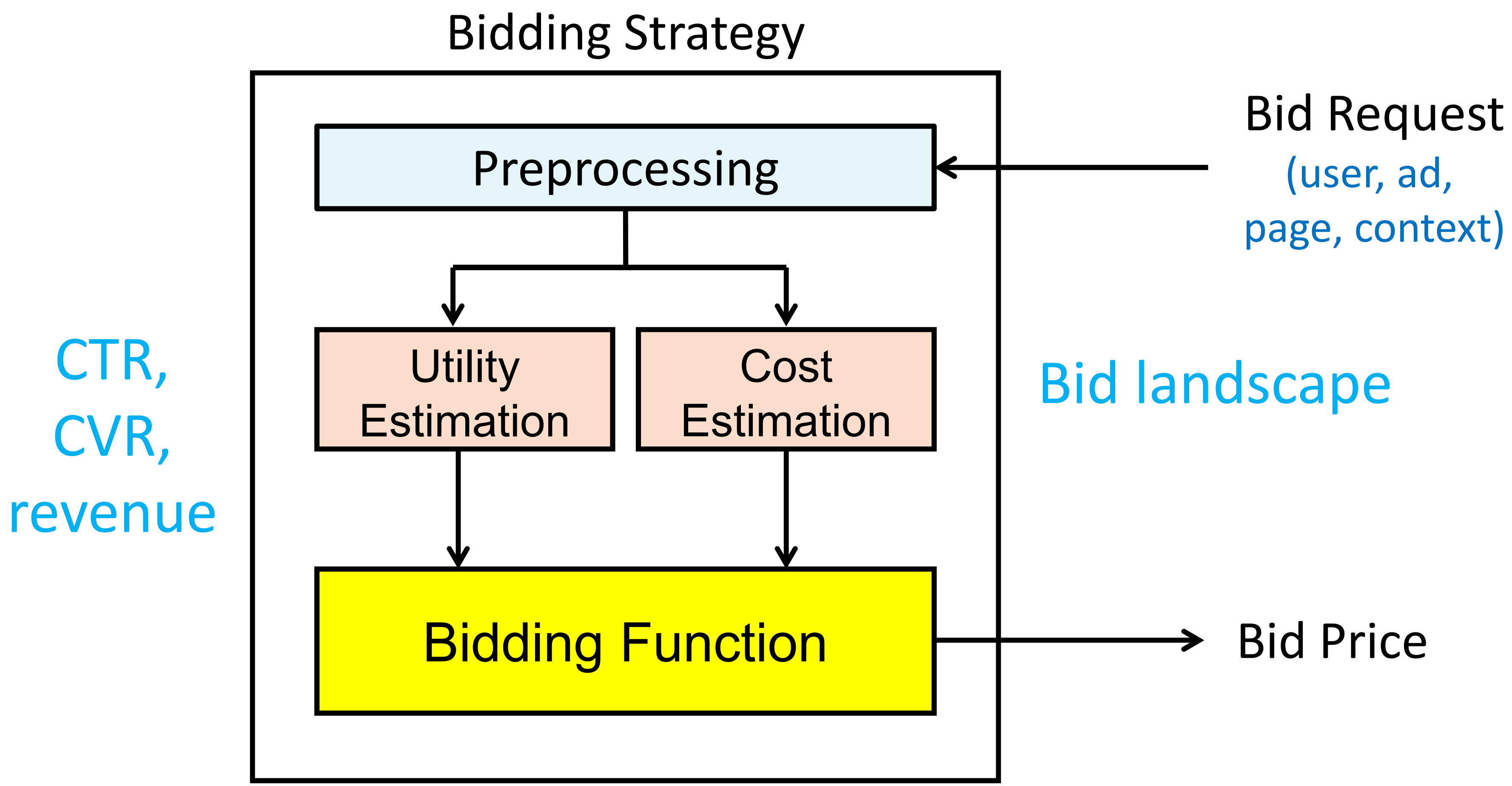
# Survival Tree Models




Node split
Based on
Clustering
categories

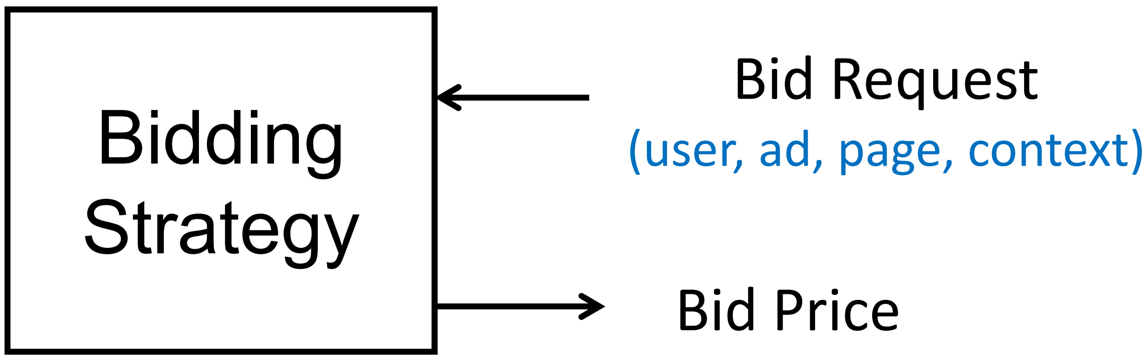
[Yuchen Wang et al. Functional Bid Landscape Forecasting for Display Advertising. ECMLPKDD 2016 ]

# Bidding Strategy in Practice:
## A Quantitative Perspective



# Bidding Strategies

- How much to bid for each bid request?




- Bid to optimize the KPI with budget constraint

$$\max_{\text{bidding strategy}} \quad \text{KPI}$$
$$\text{subject to} \quad \text{cost} \leq \text{budget}$$

# Classic Second Price Auctions

- Single item, second price (i.e. pay market price)

Reward given a bid: $$R(b) = \int_0^b (r - z)p(z)dz$$

Optimal bid: $$b^* = \max_b R(b)$$

$$\frac{\partial R(b)}{\partial b} = (r - b)p(b)$$

$$\frac{\partial R(b)}{\partial b} = 0 \Rightarrow b^* = r$$ Bid true value

# Truth-telling Bidding Strategies

- Truthful bidding in second-price auction
  - Bid the true value of the impression
  - Impression true value = $\begin{cases} \text{Value of click, if clicked} \\ \text{0, if not clicked} \end{cases}$
  - Averaged impression value = value of click * CTR
  - Truth-telling bidding:

$$\text{bid} = r_{\text{conv}} \times \text{CVR} \quad \text{or} \quad \text{bid} = r_{\text{click}} \times \text{CTR}$$

[Chen et al. Real-time bidding algorithms for performance-based display ad allocation. KDD 11]

# Truth-telling Bidding Strategies

$$\text{bid} = r_{\text{conv}} \times \text{CVR} \quad \text{or} \quad \text{bid} = r_{\text{click}} \times \text{CTR}$$

- Pros
  - Theoretic soundness
  - Easy implementation (very widely used)
- Cons
  - Not considering the constraints of
    - Campaign lifetime auction volume
    - Campaign budget
  - Case 1: $1000 budget, 1 auction
  - Case 2: $1 budget, 1000 auctions

[Chen et al. Real-time bidding algorithms for performance-based display ad allocation. KDD 11]

# Non-truthful Linear Bidding

- Non-truthful linear bidding

$$\text{bid} = \text{base\_bid} \times \frac{\text{predicted\_CTR}}{\text{base\_CTR}}$$

  - Tune base_bid parameter to maximize KPI
  - Bid landscape, campaign volume and budget indirectly considered

$$\max_{\text{bidding strategy}} \quad \text{KPI}$$
$$\text{subject to} \quad \text{cost} \leq \text{budget}$$

[Perlich et al. Bid Optimizing and Inventory Scoring in Targeted Online Advertising. KDD 12]

# ORTB Bidding Strategies

- Direct functional optimisation

winning function　　　　CTR

$$b()_{\text{ORTB}} = \underset{b()}{\arg\max} \; N_T \int_{\theta} \theta w(b(\theta)) p_\theta(\theta) d\theta$$

bidding function

$$\text{subject to} \; N_T \int_{\theta} b(\theta) w(b(\theta)) p_\theta(\theta) d\theta \le B \longleftarrow \text{budget}$$

Est. volume　　　　cost upperbound

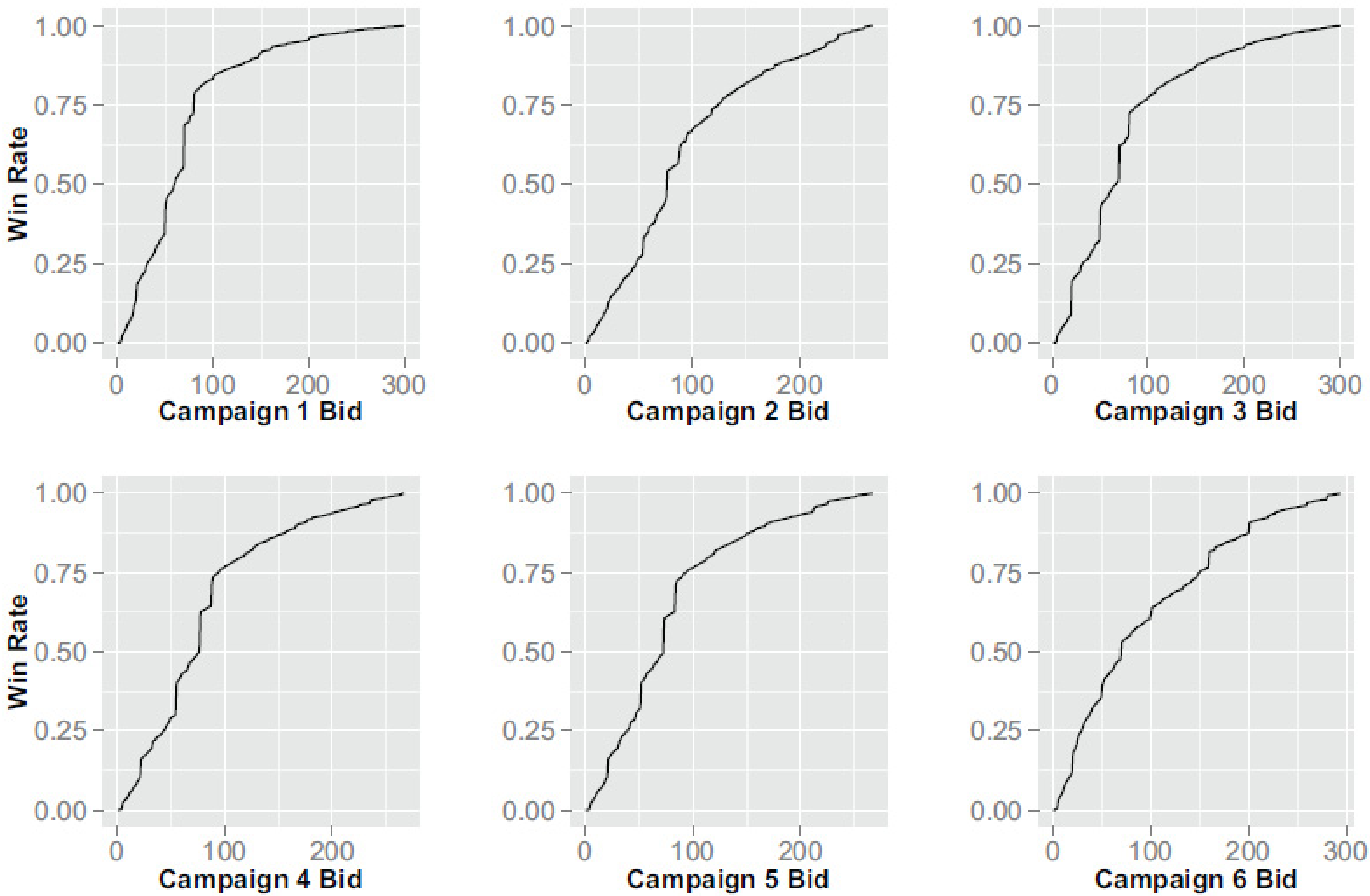
- Solution: Calculus of variations

$$\mathcal{L}(b(\theta), \lambda) = \int_{\theta} \theta w(b(\theta)) p_\theta(\theta) d\theta - \lambda \int_{\theta} b(\theta) w(b(\theta)) p_\theta(\theta) d\theta + \frac{\lambda B}{N_T}$$

$$\frac{\partial \mathcal{L}(b(\theta), \lambda)}{\partial b(\theta)} = 0 \;\; \Rightarrow \;\; \boxed{\lambda w(b(\theta)) = \Big[\theta - \lambda b(\theta)\Big] \frac{\partial w(b(\theta))}{\partial b(\theta)}}$$

[Zhang et al. Optimal real-time bidding for display advertising. KDD 14]

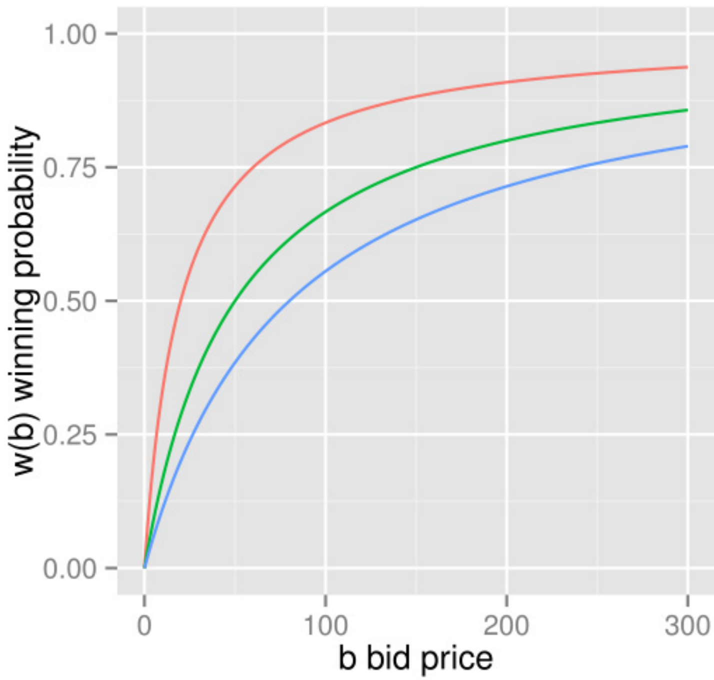
# Bid Landscape: w(bid)



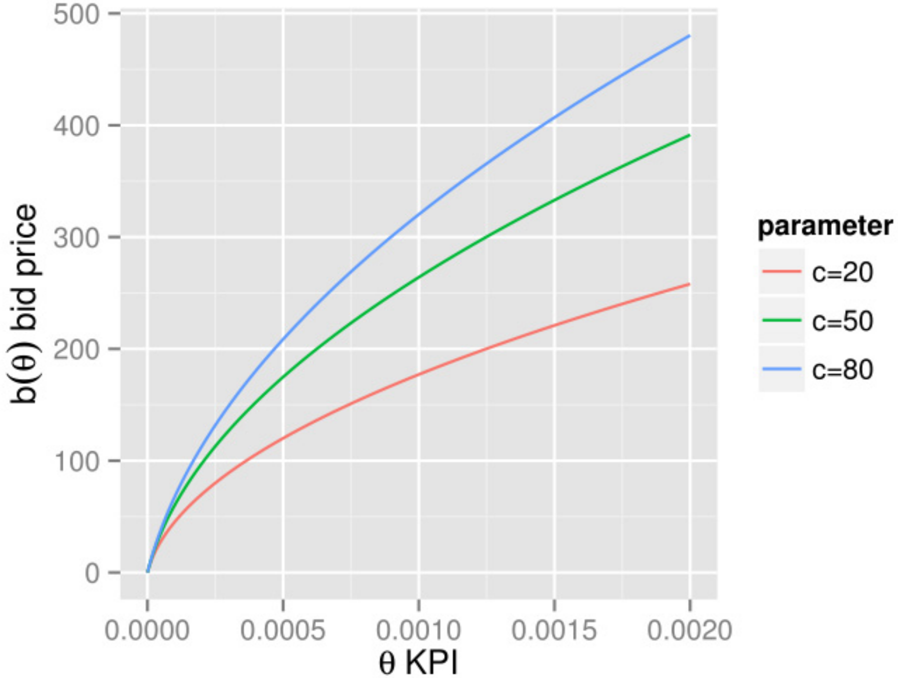

[Zhang et al. Optimal real-time bidding for display advertising. KDD 14]

# Optimal Bidding Strategy Solution

$$\lambda w(b(\theta)) = \Big[\theta - \lambda b(\theta)\Big] \frac{\partial w(b(\theta))}{\partial b(\theta)}$$




(a) Winning function 1.



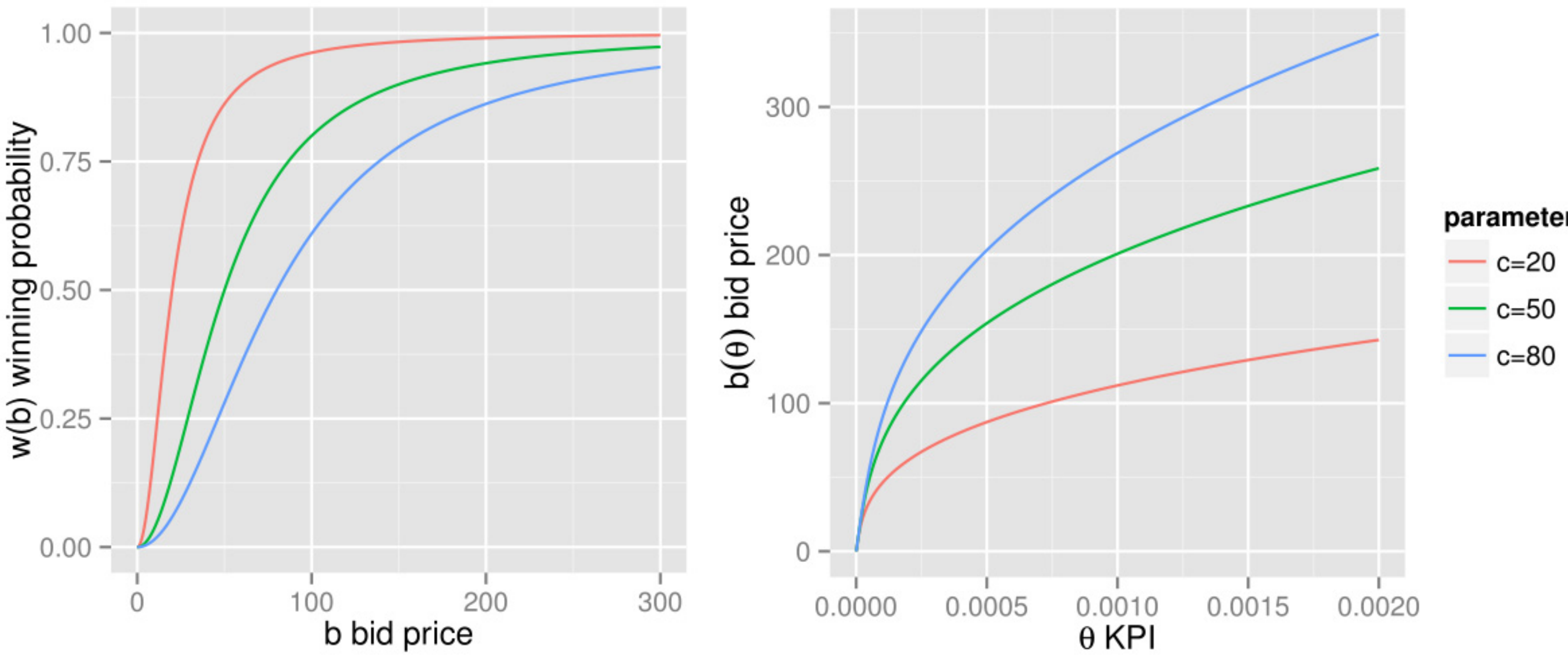

(b) Bidding function 1.

$$w(b(\theta)) = \frac{b(\theta)}{c + b(\theta)} \quad \Rightarrow \quad b_{\text{ORTB1}}(\theta) = \sqrt{\frac{c}{\lambda}\theta + c^2} - c$$

[Zhang et al. Optimal real-time bidding for display advertising. KDD 14]

# Optimal Bidding Strategy Solution

$$\lambda w(b(\theta)) = \left[\theta - \lambda b(\theta)\right] \frac{\partial w(b(\theta))}{\partial b(\theta)}$$



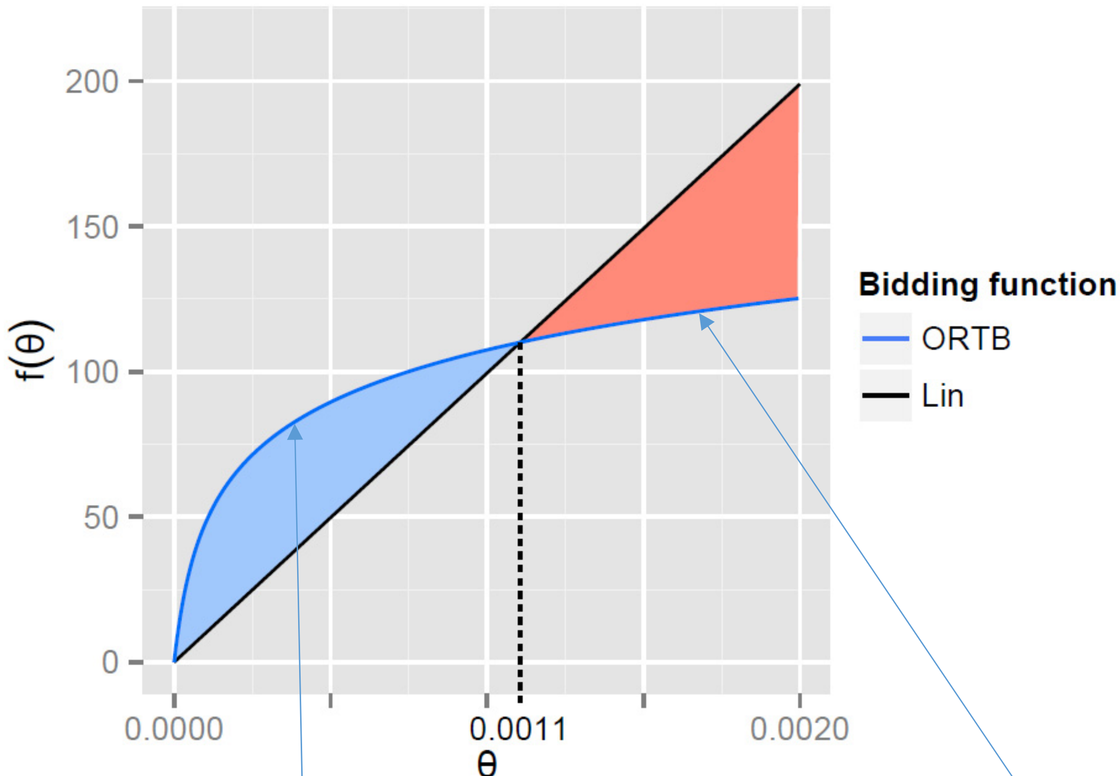

(a) Winning function 2.

(b) Bidding function 2.

$$w(b(\theta)) = \frac{b^2(\theta)}{c^2 + b^2(\theta)} \Rightarrow b_{\mathrm{ORTB2}}(\theta) = c \cdot \left[\left(\frac{\theta + \sqrt{c^2\lambda^2 + \theta^2}}{c\lambda}\right)^{\frac{1}{3}} - \left(\frac{c\lambda}{\theta + \sqrt{c^2\lambda^2 + \theta^2}}\right)^{\frac{1}{3}}\right]$$

[Zhang et al. Optimal real-time bidding for display advertising. KDD 14]

# Optimal Bidding Strategy: the Analysis




Slight increase at low bids is more effective

Thus reduce the bids at high CTR or CVR

[Zhang et al. Optimal real-time bidding for display advertising. KDD 14]

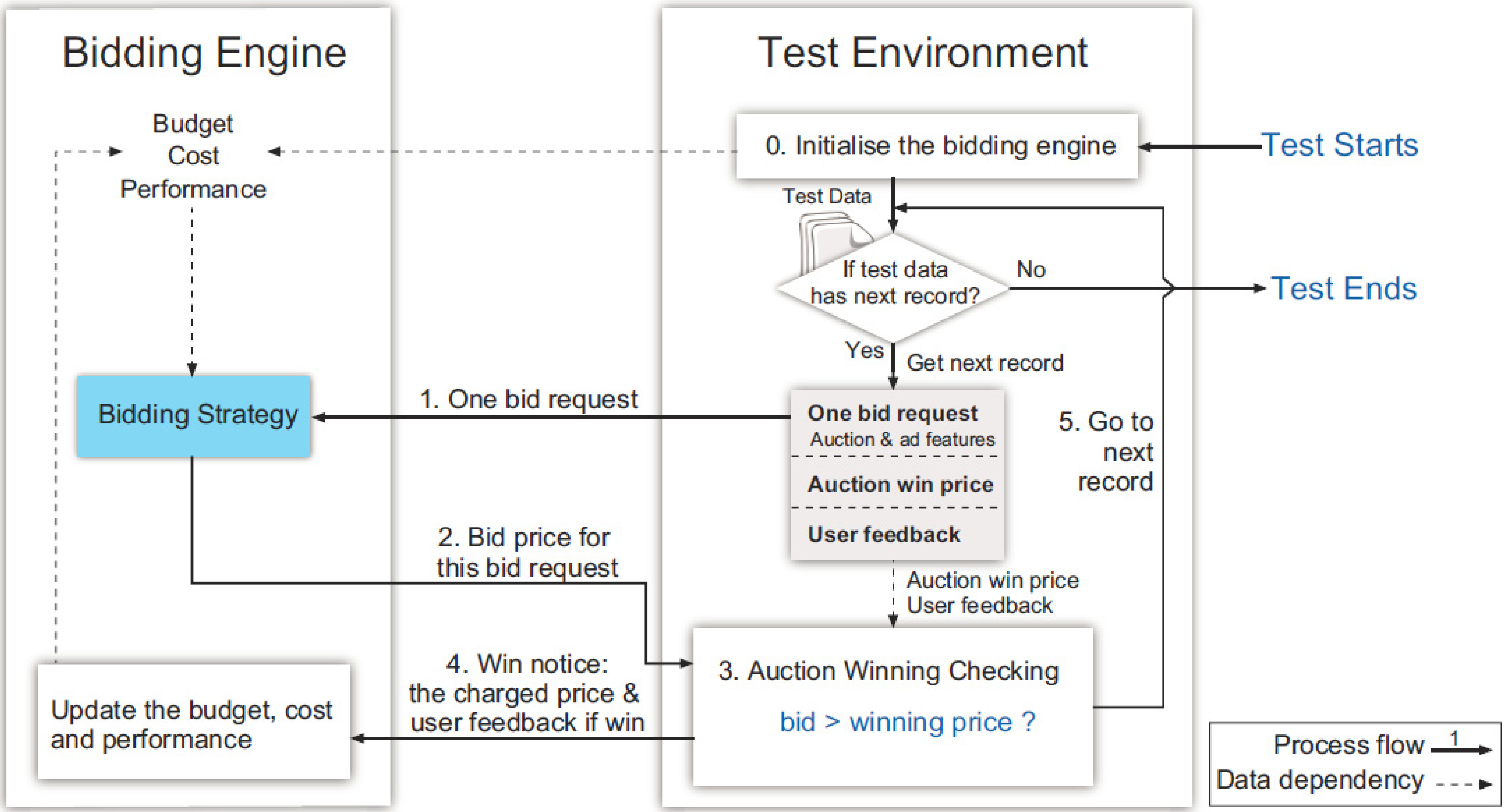
# Experiment

- We used iPinYou's dataset
  - 1-http://data.computational-advertising.org
  - 9 Campaigns,  15M impressions, 11K clicks, 935 conversions
- Evaluated bidding strategies
  - **Const**: Constant
  - **Rand**: Random
  - **Mcpc**: Bidding based on advertiser's given max eCPC [Chen et al. 2011]
  - **Lin**: Linear to pCTR [Perlich et al. 2012]
  - **ORTB1**, **ORTB2**: Optimal bidding strategies with two forms of winning rate functions



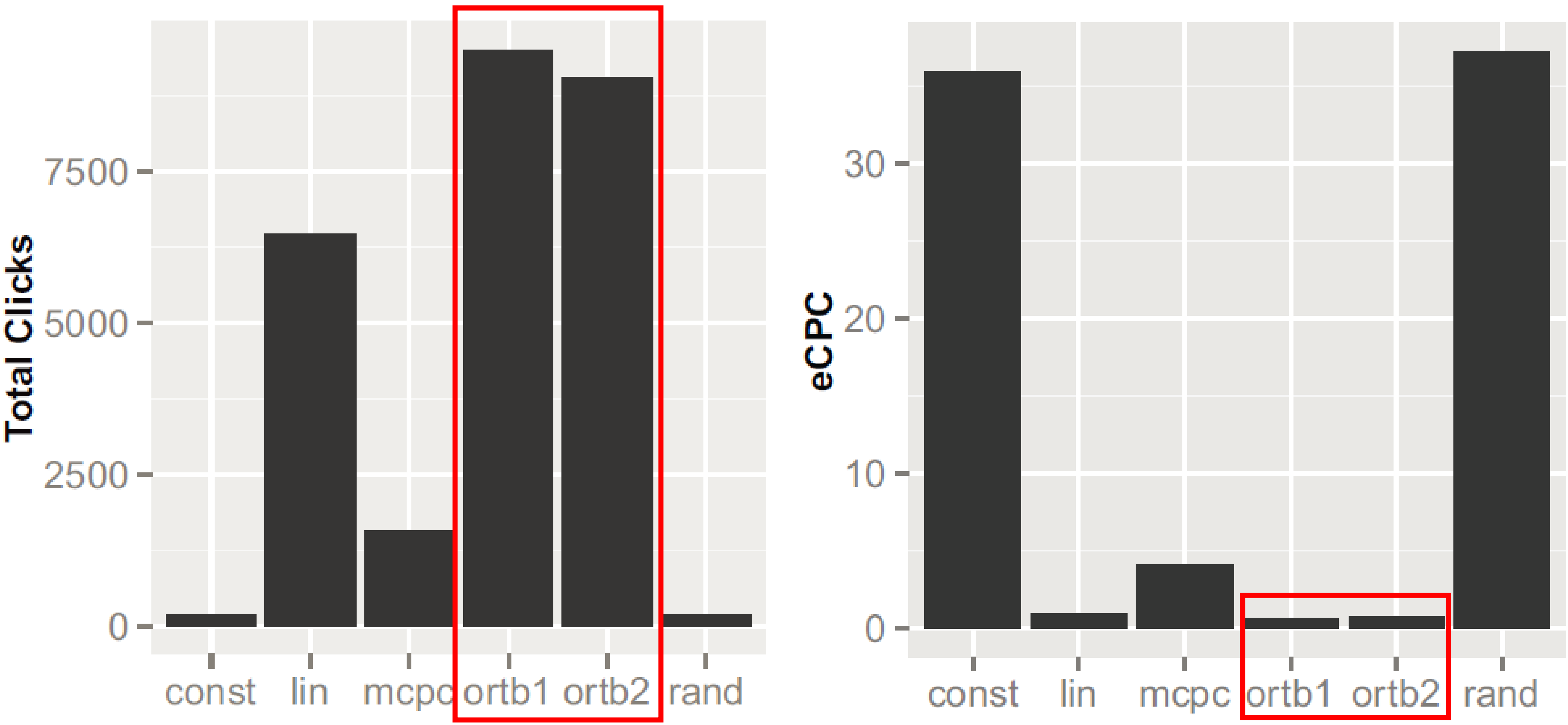
[Zhang et al. Optimal real-time bidding for display advertising. KDD 14]

# Offline Test Evaluation Flow




[Zhang et al. Optimal real-time bidding for display advertising. KDD 14]

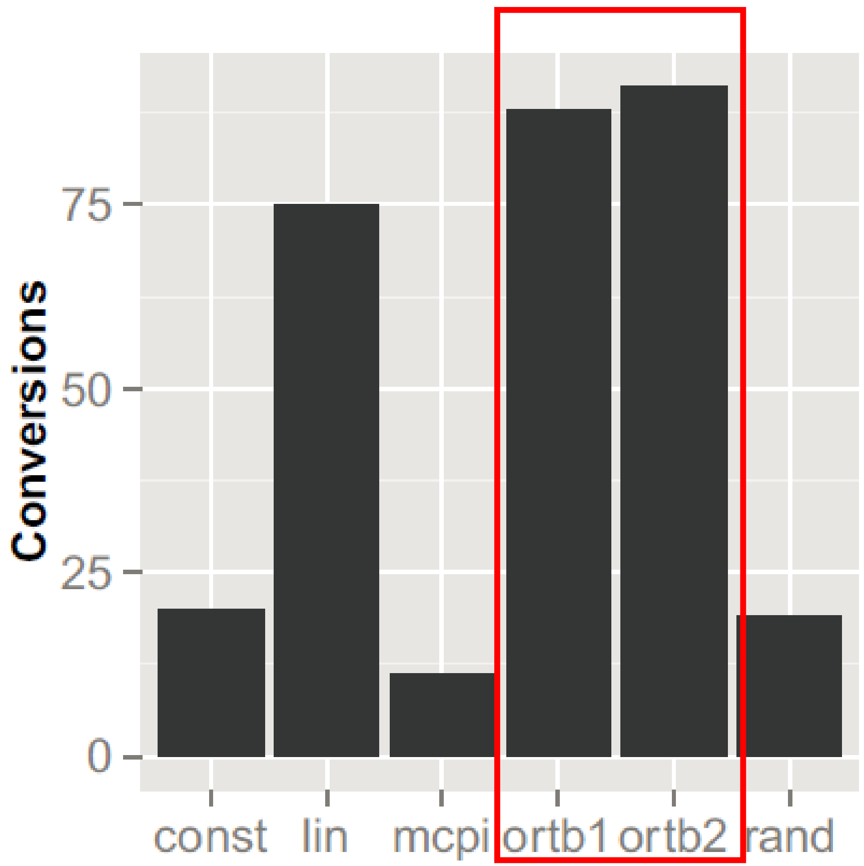
# Overall performance: Optimizing Clicks






[Zhang et al. Optimal real-time bidding for display advertising. KDD 14]

# Overall performance – Optimizing Conversions






[Zhang et al. Optimal real-time bidding for display advertising. KDD 14]

# Unbiased Optimization

- Bid optimization on 'true' distribution

$$\arg\max_{b()} \; T \int_{\boldsymbol{x}} f(\boldsymbol{x}) w(b(f(\boldsymbol{x}))) p_x(\boldsymbol{x}) d\boldsymbol{x}$$

$$\text{subject to} \;\; T \int_{\boldsymbol{x}} b(f(\boldsymbol{x})) w(b(f(\boldsymbol{x}))) p_x(\boldsymbol{x}) d\boldsymbol{x} = B$$

- Unbiased bid optimization on biased distribution

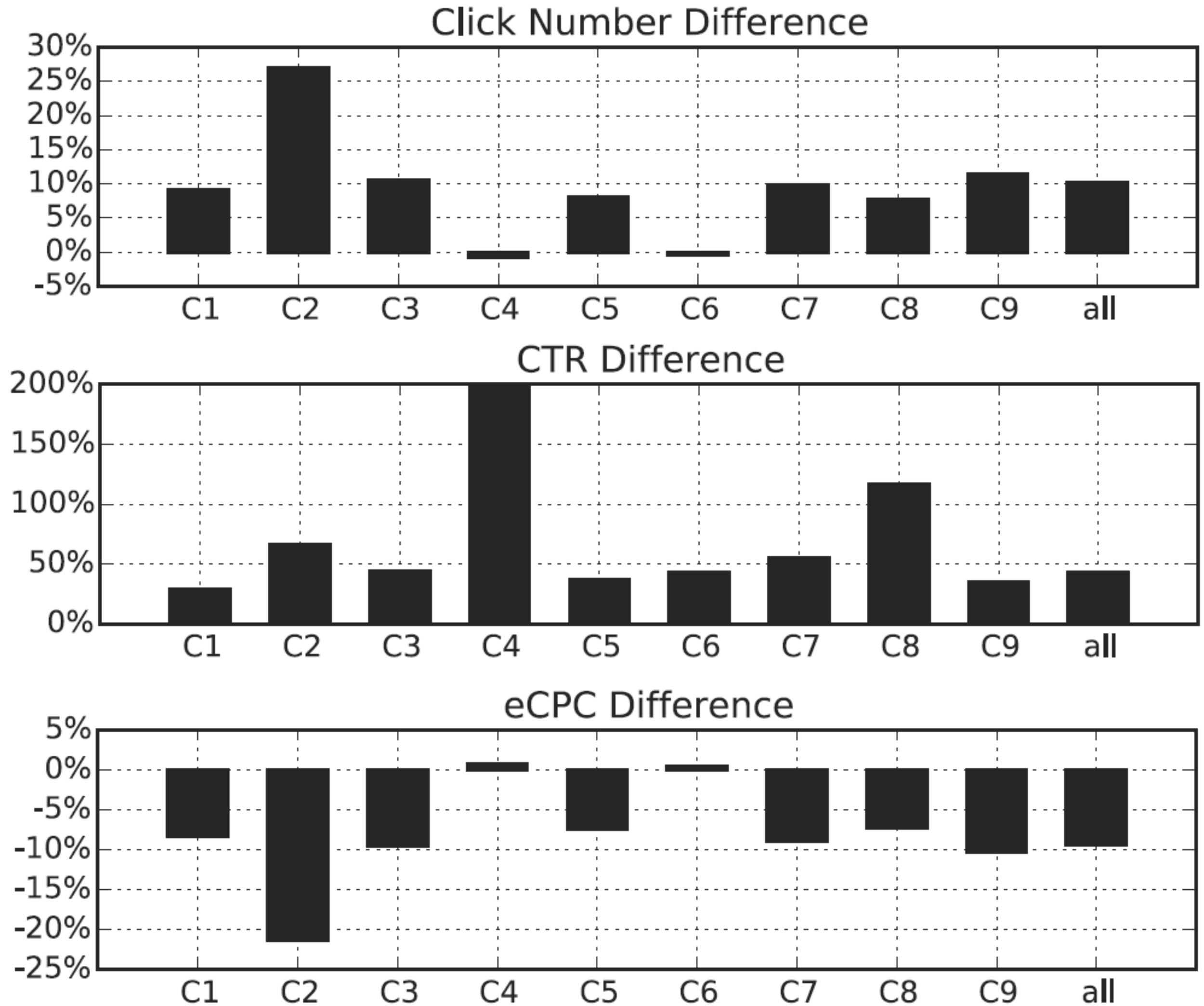
$$\arg\max_{b()} \; T \int_{\boldsymbol{x}} f(\boldsymbol{x}) w(b(f(\boldsymbol{x}))) \frac{q_x(\boldsymbol{x})}{w(b_{\boldsymbol{x}})} d\boldsymbol{x}$$

$$\text{subject to} \;\; T \int_{\boldsymbol{x}} b(f(\boldsymbol{x})) w(b(f(\boldsymbol{x}))) \frac{q_x(\boldsymbol{x})}{w(b_{\boldsymbol{x}})} d\boldsymbol{x} = B$$

[Zhang et al. Bid-aware Gradient Descent for Unbiased Learning with Censored Data in Display Advertising. KDD 2016.]

# Unbiased Bid Optimization

A/B Testing on Yahoo! DSP.




[Zhang et al. Bid-aware Gradient Descent for Unbiased Learning with Censored Data in Display Advertising. KDD 2016.]

# Content of This Course

- Real-time bidding based display advertising
- User tracking and profiling
- Real-time bidding strategies
- Fraud detection

# Fraud

- Reported by Interactive Advertising Bureau’s (IAB) in 2015

- Ad fraud is costing the U.S. marketing and media industry an estimated $8.2 billion each year

- $4.6 billion, or 56%, of the cost to “invalid traffic”, of which 70% is performance based, e.g., CPC and CPA, and 30% is CPM based.

Interactive Advertising Bureau. What is an untrustworthy supply chain costing the us digital advertising industry?, 2015.

# An Display Ad Example

How do you know the user is a human or a robot?

## 大陆

**河南省公安厅彻查“封丘36人入警35人身份不合规”**

中封丘县公安局的36名受训人员，35人是公安局内部的文职或临时人员，与“民警必须具备公务员身份”的国家规定不符，引发该局内部

- 上海至成都沿江高铁提上日程 串联长江沿线22城市
- 2016号歼-20原型机曝光 已滑行测试(图)
- 日媒：中国或派万吨海警船巡钓鱼岛 打消耗战
- 外媒：中国开始研制隐身武装直升机 预计2020年交付
- 习近平关于中美关系的十个判断
- 住建部黑臭水沟整治工作指南：9成百姓满意才能达标
- 陕西：职校“校长”让女学生陪酒 学校被撤除
- 揭秘“团团伙伙”的武钢漩涡和落马高管

## 国际

**巴塞罗那200万人游行 呼吁加泰罗尼亚独立(图)**

- 李炜光：收税是不公平的恶？
- 许章润：超级大国没有纯粹内政
- 刘昀献:国外政党联系群众的路径研究

### 时局观

民革中央副主席：中共从未否定国民党抗战作用

- 施芝鸿：文革基础上搞改革致一个时期市场官场乱象
- 朱维群回应争议：尊重民族差异而不强化
- 伊协副会长：穆斯林不应因宗教功修忽视社会责任

### 领袖圈

奥巴马54岁啦，当7年总统人苍老了头发也



海绵城市 未来之城
水危机：青岛告急
探访中国绿化博览会
帝都吸引华人首富
凤凰房产 诚邀加盟
谈华山论剑与中国精神
黑龙江创新驱动三步棋
《印记》之江城夜未眠
办公环境搜查令
圈层生活尽在凤凰会

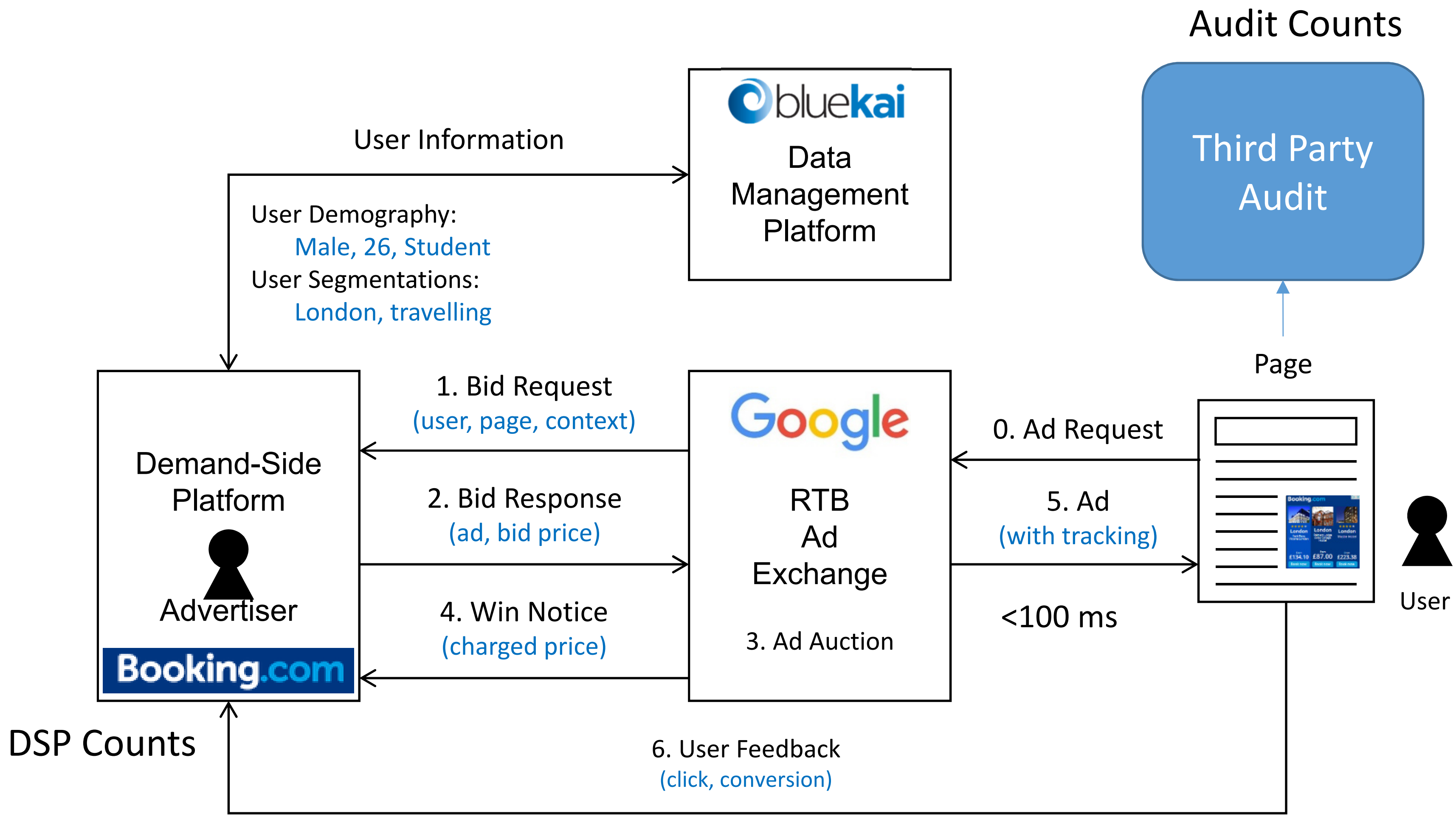
### 精彩视频

凤凰联播台

菲媒曝菲律宾军演针对中国 直指南海生命线

播放数：2602282

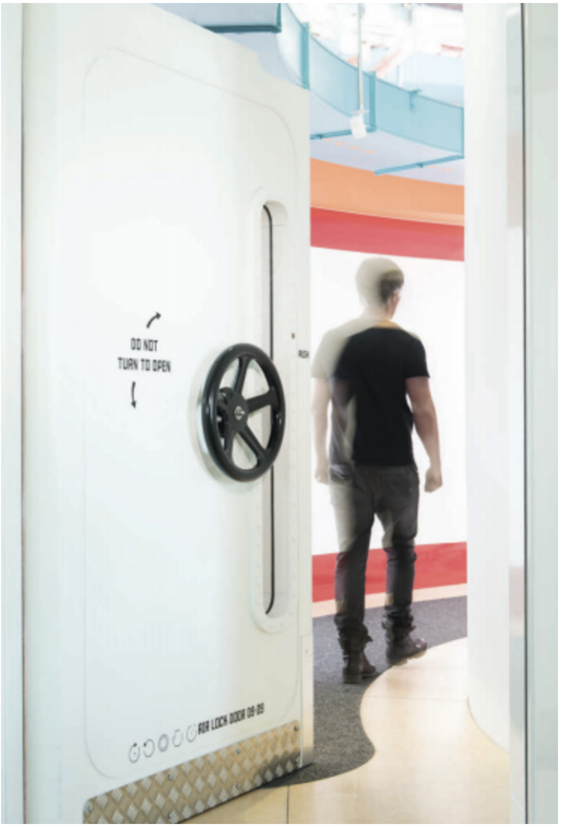
# Leverage Third Party to Audit




- Typically, the counts of the DSP and Audit should be close
  - Say $\pm 5\%$

# A Good Story of Fraud Fighters

- http://www.rtbchina.com/inside-google-s-secret-war-ad-fraud.html



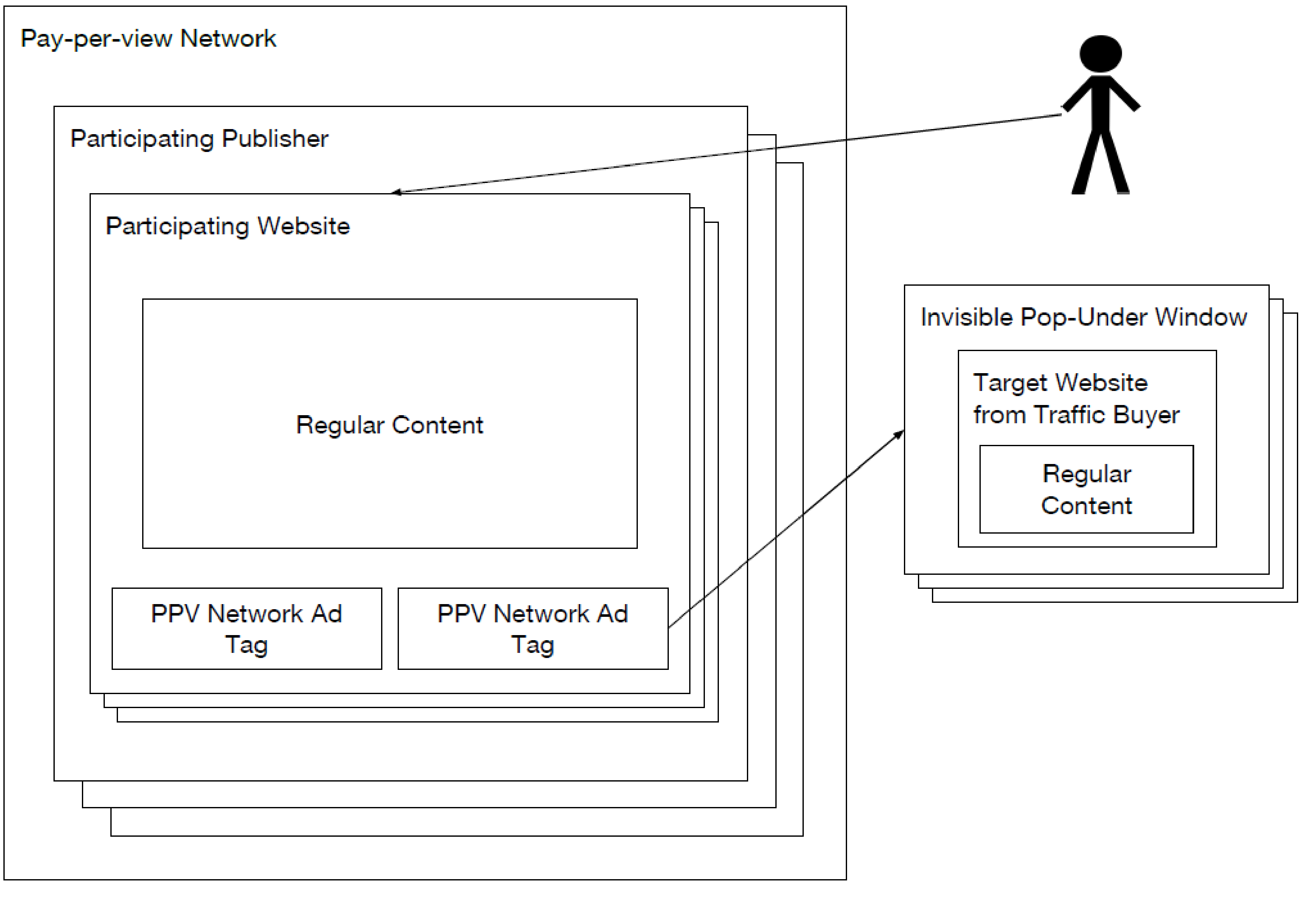

# Ad Fraud Types

- Impression fraud
  - where the fraudster generates fake bid requests, sells them in ad exchanges, and gets paid when advertisers buy them to get impressions
- Click fraud
  - where the fraudster generates fake clicks after loading an ad
- Conversion fraud
  - where the fraudster completes some actions, e.g., filling out a form, downloading and installing an app, after loading an ad

# Ad Fraud Sources

- Publisher driven: pay-per-view network
- User/robot driven: botnet

# Pay-Per-View (PPV) Networks



# Possible Methods to Avoid PPV for Advertisers

- Viewport size check: valid impressions will not be displayed in a 0x0 viewport, which is invisible to users

- A referrer blacklist, which checks if the traffic is from the PPV networks

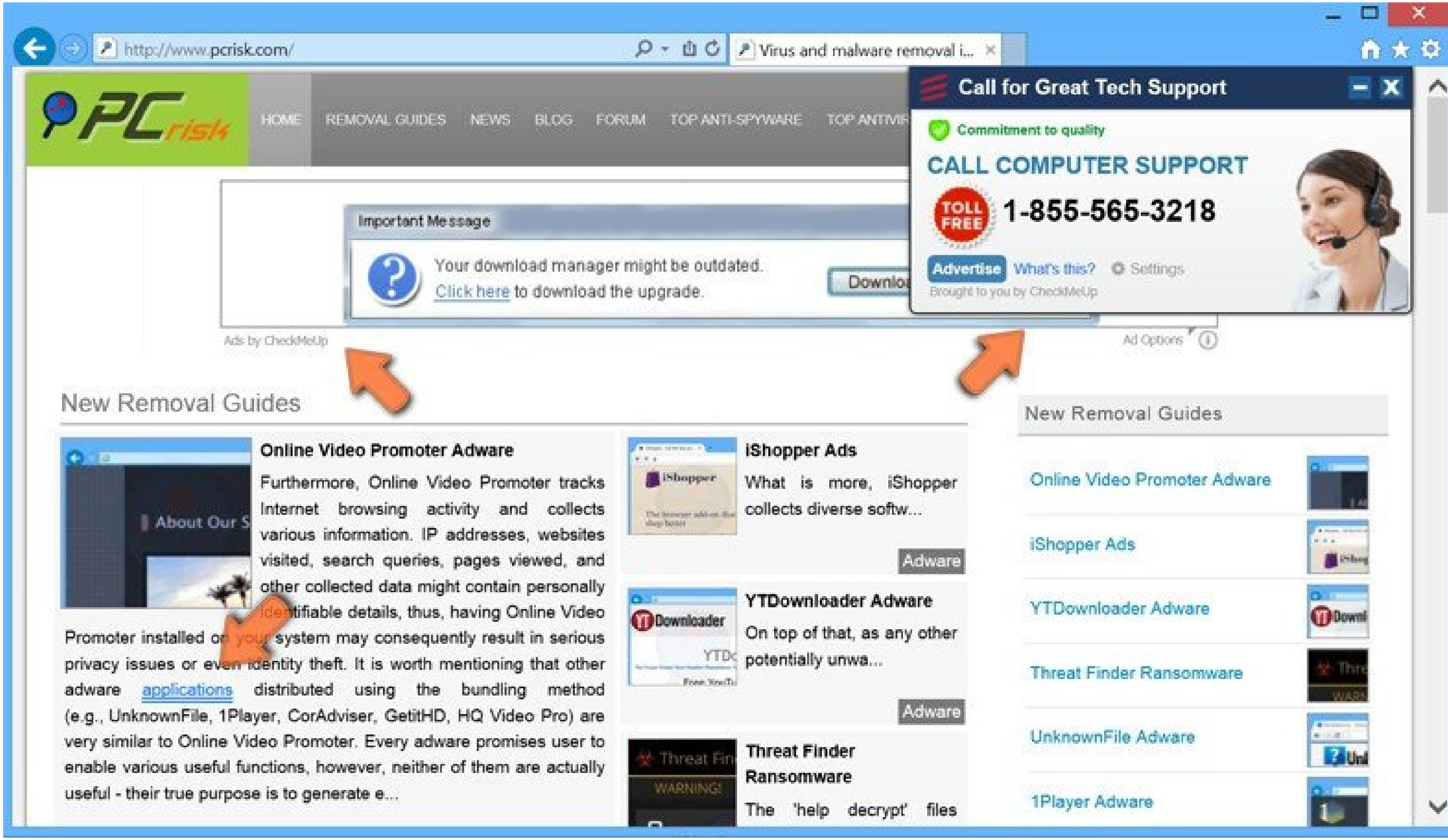
- A publisher blacklist, which avoids buying traffic from publishers who participate in the PPV networks

# Botnets

- Botnets are usually built with compromised end users’ computers.
  - These computers are installed with one or multiple software packages, which run autonomously and automatically.
  - Adware

BotnetsMaryam Feily, Alireza Shahrestani, and Sureswaran Ramadass. A survey of botnet and botnet detection. In *2009 Third International Conference on Emerging Security Information, Systems and Technologies*, pages 268–273. IEEE, 2009.

# Adware Examples



# A Few Ways to Detecting Botnets

- Signature based detection, which extracts software / network package signature from known botnet activities
- Anomaly detection of traffic
- DNS based detection, which focuses on analyzing DNS traffic which is generated by communication of bots and the controller
- Mining based detection, which uses Machine Learning techniques to cluster or classify botnet traffic

# Data Mining based Fraud Detection

- Ad fraud detection is usually an unsupervised learning problem and it is difficult to capture the ground-truth

- Fully unsupervised learning
  - Detect the fraud based on the revealed web structures and human heuristics

- Semi-supervised learning
  - Detect the fraud by training a predictor based on a very small labeled data and large unlabeled data

# Ad Fraud Detection with Co-visit Networks

- Define a bipartite graph between users (browsers) and websites

G = <*B*, *W*, *E*>

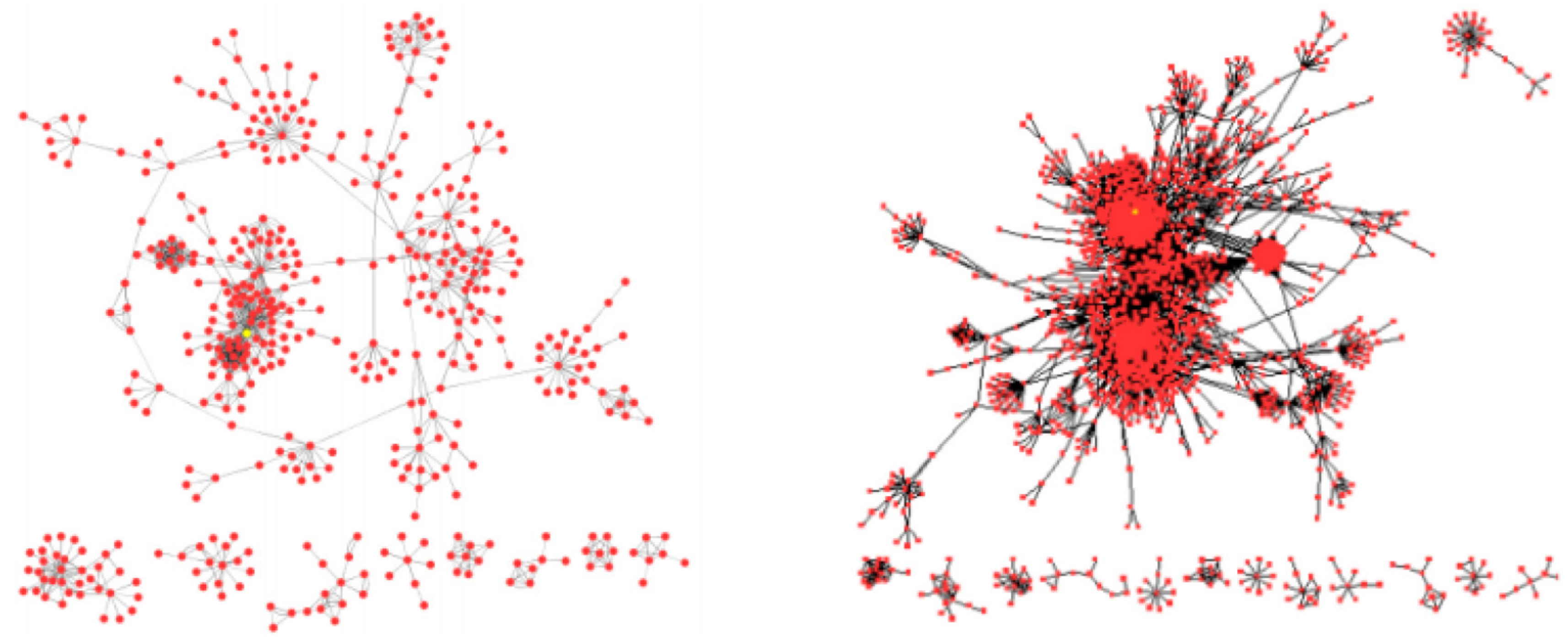
  - *B*: users
  - *W*: websites
  - *E*: the edge indicating whether the user has visit the website over a specified time period

- The co-visit network is based on *G*

$$G_W^n = \langle V_W \subseteq W, E = (x, y) : x, y \in W, [\Gamma_G(x) \cap \Gamma_G(y)]/\Gamma_G(x) \geq n \rangle$$

Ori Stitelman. Using co-visitation networks for detecting large scale online display advertising exchange fraud.KDD 2013.

# Co-Visit Network Examples



- The co-visit networks of Dec 2010 (left) and Dec 2011 (right) reported by Stitelman et al. [2013].

Ori Stitelman. Using co-visitation networks for detecting large scale online display advertising exchange fraud.KDD 2013.

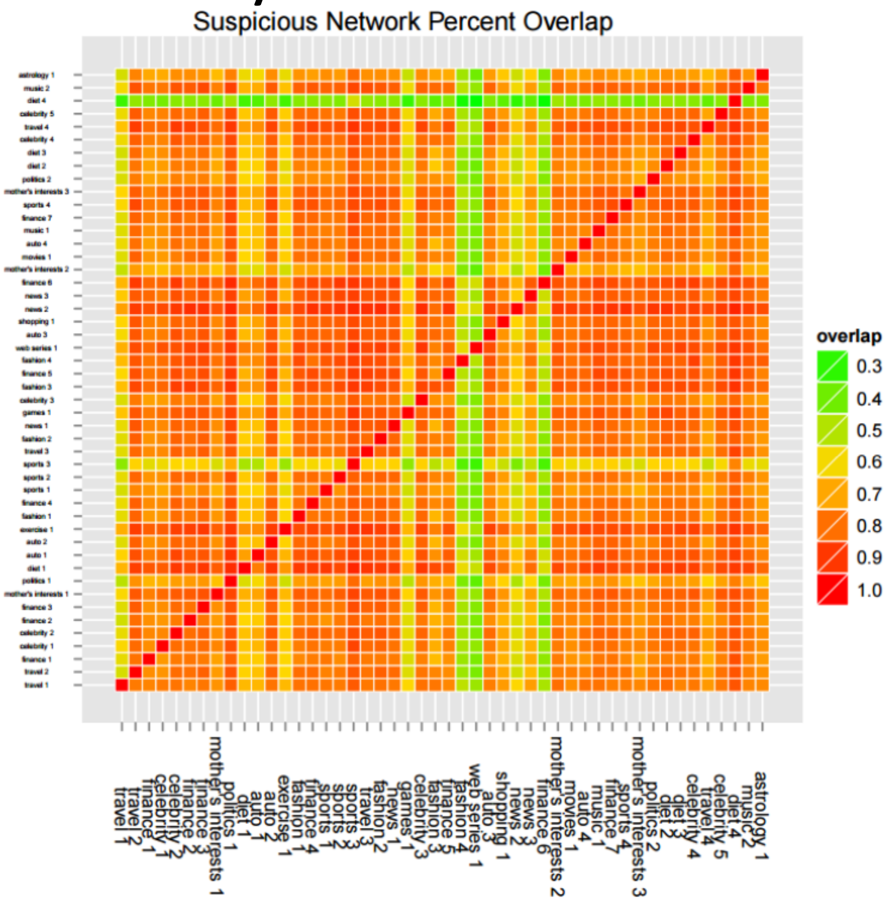
# Co-Visit Network for Fraud Detection

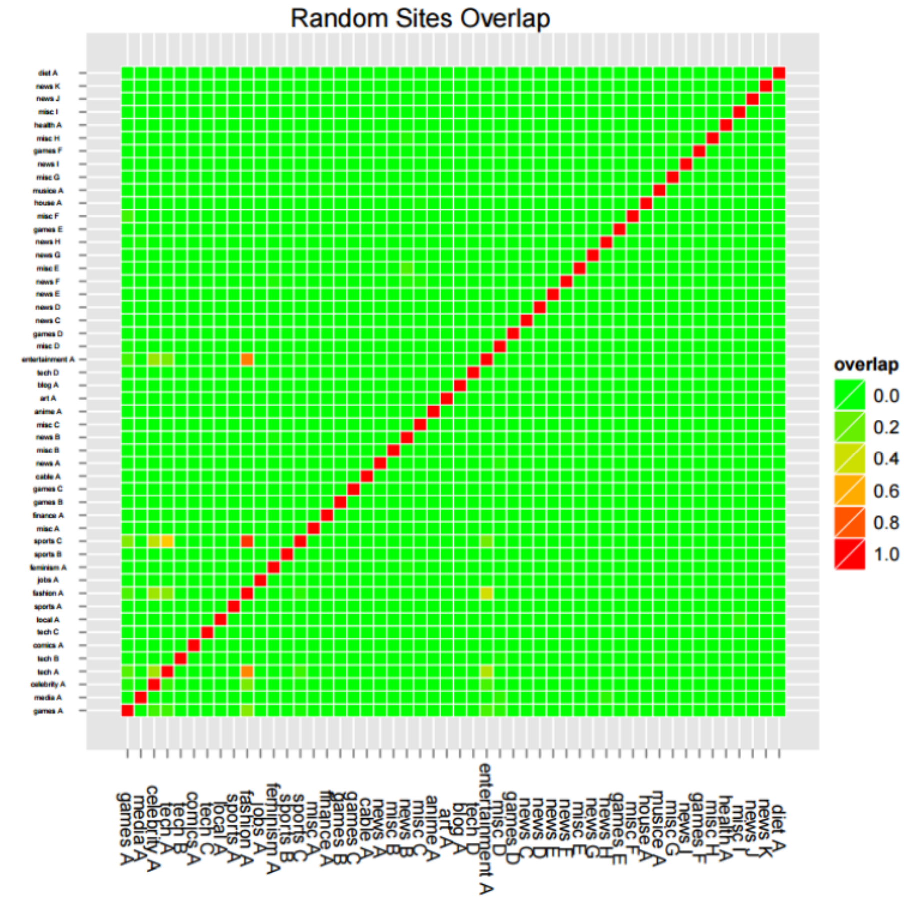
- Intuition: two websites' user overlap is normally very small
    - High dimensional random vectors are almost vertical (i.e. with cosine close to 0)

Ori Stitelman. Using co-visitation networks for detecting large scale online display advertising exchange fraud.KDD 2013.

# Co-Visit Network for Fraud Detection

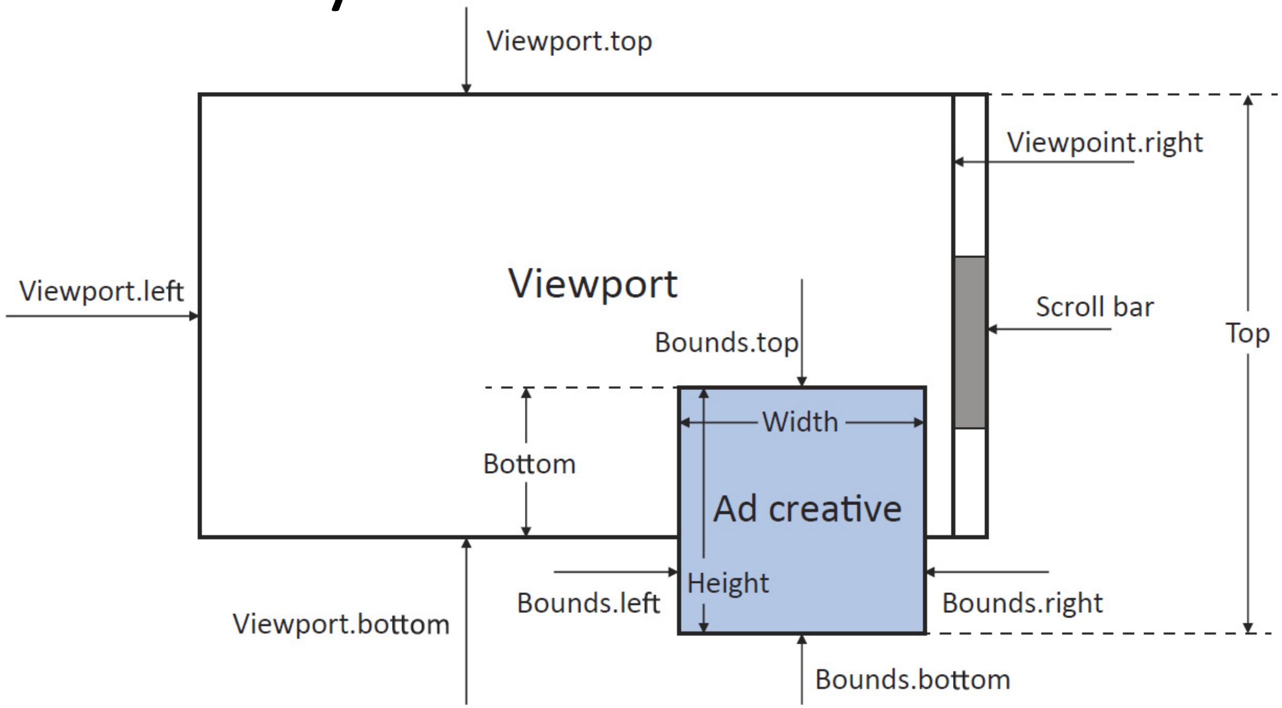
- Intuition: two websites' user overlap is normally very small




Ori Stitelman. Using co-visitation networks for detecting large scale online display advertising exchange fraud.KDD 2013.

# Viewability Methods

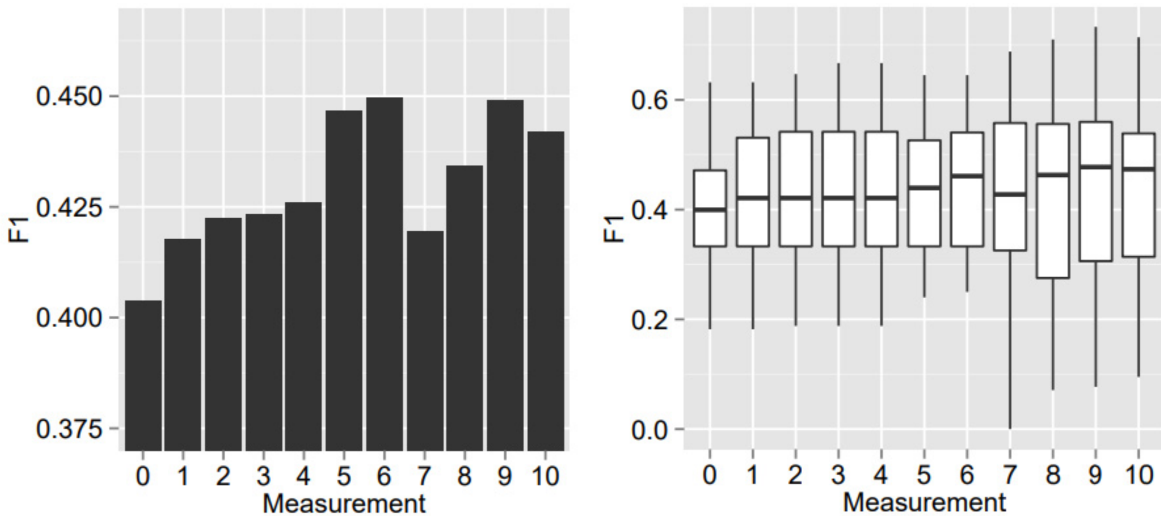



We developed a javascript to tracking each user's behavior on browsing a displayed ad

- Pixel percentage tracking: The displayed pixel percentage for rectangle ad creative in the viewport
- Exposure time tracking: The exposure time is associated with a pixel percentage threshold.

Weinan Zhang, Ye Pan, Tianxiong Zhou, and Jun Wang. An empirical study on display ad impression viewability measurements. *arXiv* 2015.

# Viewability Methods




| meas. | pixel | time | meas. | pixel | time | meas. | pixel | time |
|---|---|---|---|---|---|---|---|---|
| 0 | 0% | 0 | 4 | 100% | (0,2) | 8 | 50% | [4,+) |
| 1 | 0% | (0,2) | 5 | 50% | [2,4) | 9 | 75% | [4,+) |
| 2 | 50% | (0,2) | 6 | 75% | [2,4) | 10 | 100% | [4,+) |
| 3 | 75% | (0,2) | 7 | 100% | [2,4) | | | |

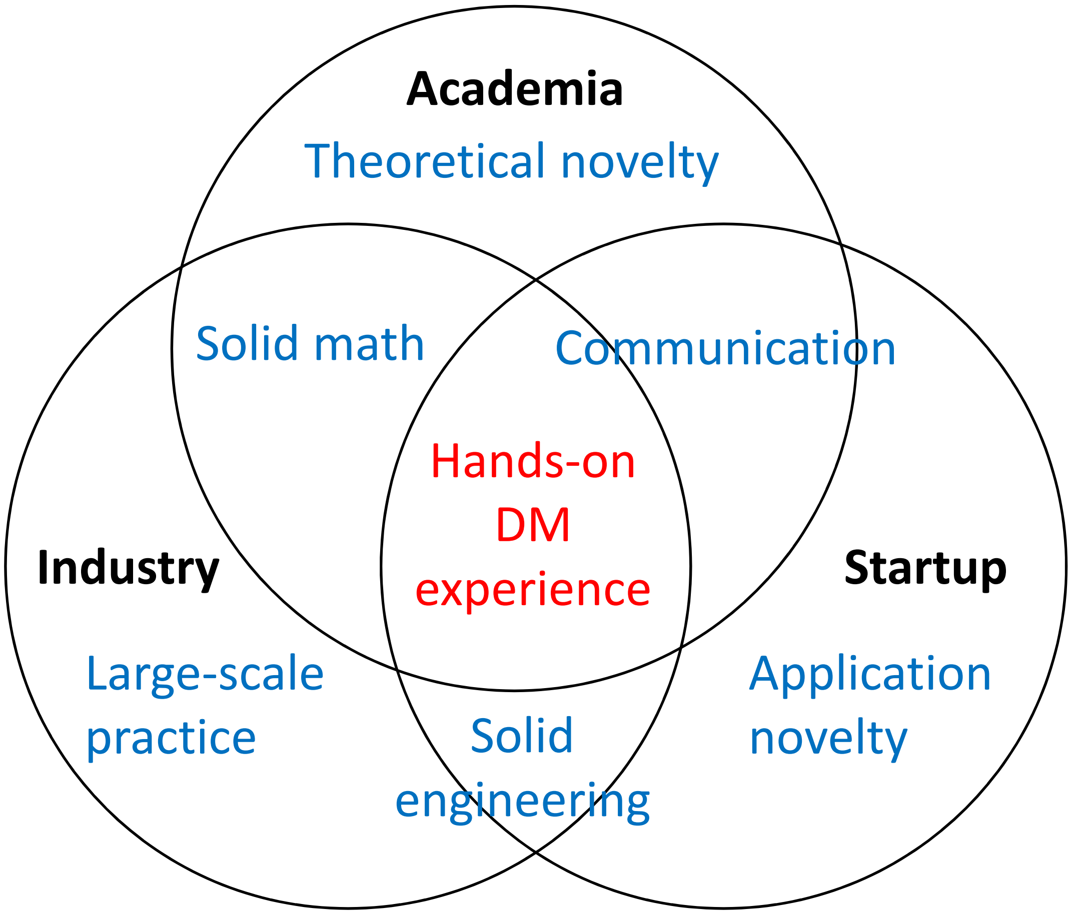
- Results: (pixel ≥ 75%, time ≥ 2s) provided the highest average F1 score and median F1 score

Weinan Zhang, Ye Pan, Tianxiong Zhou, and Jun Wang. An empirical study on display ad impression viewability measurements. *arXiv* 2015.

# Summary of EE448

1. Data Mining Intro
2. Fundamentals of Data
3. Basic DM Algorithms
4. Supervised Learning 1
5. Supervised Learning 2
6. Supervised Learning 3
7. Supervised Learning 4
8. Unsupervised Learning
9. Search Engines
10. Ranking Information Items
11. Recommender Systems
12. Computational Ads
13. Behavioral Targeting
14. Poster Session

# We focus on hands-on DM




- Get familiar with various data mining applications.
- Play with the data and get your hands dirty!

# Thank You!

Weinan Zhang, Ph.D.
Assistant Professor

John Hopcroft Center for Computer Science
Dept. of Computer Science & Engineering
Shanghai Jiao Tong University